Symmetric Causal Independence Models

Rasa Jurgelėnaitė

SIKS Dissertation Series No. 2009 - 01
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

This research has been funded by the Netherlands Organisation for Scientific Research (NWO) under project number FN4556 and a Vici grant (639.023.604) awarded to Tom Heskes.

ISBN: 978-90-9023645-2

Cover design: Dovilė Jurgelėnaitė

Symmetric Causal Independence Models

Een wetenschappelijke proeve op het gebied van de
Natuurwetenschappen, Wiskunde en Informatica

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. mr. S.C.J.J. Kortmann,
volgens besluit van het College van Decanen
in het openbaar te verdedigen op maandag 19 januari 2009
om 15.30 uur precies

door

Rasa Jurgelėnaitė
geboren op 28 april 1979
te Kaunas

Promotor: Prof. dr. T.M. Heskes

Manuscriptcommisie:
Prof. dr. H.J. Kappen
Prof. dr. A.P.J.M. Siebes (Universiteit Utrecht)
Dr. J.M. Peña (Linköping University)

Acknowledgments

I would like to start this thesis by expressing my gratitude to everyone who one way or another contributed to this thesis being written or to my decision to become a PhD student in machine learning.

First of all, I would like to thank my supervisor Tom Heskes. His knowledge, insights, enthusiasm and dedication do not stop surprising me. Tom, you are a great supervisor, I learned a lot from you.

I would also like to thank Tjeerd Dijkstra and Clemens Kocken, the co-authors of the paper on gene regulation in the malaria parasite, from whom I learned so many interesting things about genomics and malaria. I am also grateful to Theo van der Weide on whom I could always rely in difficult times and Peter Lucas who gave me an opportunity to become a Phd student.

I am grateful to my colleagues who agreed to read and comment on my thesis. Marcel, Perry, Adriana, Botond and Henriëtte, thank you for finding time to improve my thesis. I also want to thank my dear friend Nelė Šimoliūnienė, who read the Lithuanian summary of my thesis and made it less English and more Lithuanian.

I would also like to thank a group of people, who, in my opinion, are too often undeservedly forgotten in the acknowledgments section. I was lucky to have so many dedicated teachers and supervisors who not only taught me so much about their subjects, but also were great examples to me. I would especially like to thank three of them: my mathematics teacher at the gymnasium Zenonas Repčys, my bachelor's thesis supervisor Gailius Raškinis and my master's thesis supervisor José M. Peña.

I had some really good times with my colleagues, who are amazing in so many different ways. Guys, I will miss our drinks together, game evenings

and salsa lessons. Marcel, I would like to thank you for your friendship and laughs we had together. Stefan and Theodore, you two made the TimeBayes project a lot of fun.

I was lucky to get two dear friends in the Netherlands: Deimantė Stijvers and Natasha Stash. Thank you, girls, for being there for me.

My warmest thanks go to my family. I would not be writing this thesis if it was not for my parents, who have taught me to love books and to respect knowledge and education, who worked very hard so I and my sister would not miss anything. I want to thank them and my sister Dovilė (the designer of the beautiful cover of this thesis) for their continuous support, patience and understanding - it is everything but easy when family members live so far away from each other.

My gratitude also goes to Wouter's family who are so incredibly kind to me and always treat me as a member of their family.

Finally, I would like to thank my dear Wouter who is always a rock for me to lean against. Schatje, we had to go far away, all the way to Denmark, to find each other. It would have been worth to go even further... I am looking forward to the times ahead of us.

Contents

Chapter 1

Introduction

This thesis studies the problems of probabilistic inference and parameter learning in symmetric causal independence models and demonstrates that symmetric causal independence models can be successfully applied in medical and genomic domains. In this chapter, we present relevant basic concepts and general knowledge about the application domains.

1.1 Bayesian networks

A *Bayesian network* [85] provides a graphical representation of a probability distribution over a set of random variables; it consists of nodes that represent the variables and arcs that encode conditional independencies between the variables.

To motivate the use of directed graphs to describe probability distributions, we need to define the *conditional independence* between two sets of random variables. Let $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ denote sets of random variables. The variables in $\mathbf{X}$ are said to be conditionally independent of $\mathbf{Y}$ given $\mathbf{Z}$, if the following condition holds:

$$\Pr(\mathbf{X} \mid \mathbf{Y}, \mathbf{Z}) = \Pr(\mathbf{X} \mid \mathbf{Z}).$$

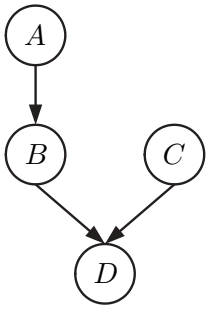


Figure 1.1: A Bayesian network representing the joint probability distribution over the variables A, B, C and D.

Exploiting conditional independencies allows us to express the joint probability in terms of the local distribution and reduce the number of parameters to be estimated:

$$\Pr(X_1, \ldots, X_n) = \prod_{i=1}^{n} \Pr(X_i \mid \pi(X_i)),$$

where $\pi(X_i)$ denotes the set of parents of X_i. Variable X_j is a parent of variable X_i if there is an arc going from variable X_j to variable X_i.

As an illustration, consider the example Bayesian network shown in Figure 1.1. By using the chain rule of probability, the joint probability distribution over the variables in this Bayesian network can be written as

$$\Pr(A, B, C, D) = \Pr(A) \Pr(B \mid A) \Pr(C \mid A, B) \Pr(D \mid A, B, C).$$

By using conditional independencies, the joint probability distribution becomes

$$\Pr(A, B, C, D) = \Pr(A) \Pr(B \mid A) \Pr(C) \Pr(D \mid B, C).$$

Thus, we can see that conditional independencies allow us to represent the joint probability more compactly. The savings are much more pronounced in larger Bayesian networks.

Next to providing a framework for deriving efficient probabilistic inference and learning algorithms, Bayesian networks have a number of other useful properties. Firstly, they provide an intuitive representation of domain

knowledge. Secondly, Bayesian networks allow combining expert knowledge and statistical data. Finally, Bayesian networks are well suited to deal with incomplete data.

1.2 Probabilistic inference

Probabilistic inference in Bayesian networks is the task of computing conditional probabilities of the values of some of the nodes (*hidden* or *unobserved* nodes) given the values of other nodes (*evidence* or *observed* nodes). The problem of probabilistic inference was shown to be NP-hard [20]. To make the problem more tractable, researchers proposed a number of *exact* and *approximate* inference algorithms.

Exact inference algorithms exploit conditional independencies in the joint probability distribution. The most widely-used exact inference methods are *clique tree propagation* [54, 68, 101] and *variable elimination* [25, 127].

Clique tree propagation (CTP) is based on a secondary structure called a clique tree (also known as a junction tree or join tree) which is built in three steps. Firstly, the moral graph of a Bayesian network is constructed by connecting non-connected parent variables that share a common child variable and removing the directionality of the arcs. Secondly, a triangulated graph is formed by adding arcs selectively to the moral graph so that any two non-adjacent vertices on a cycle would have an edge connecting them. Finally, the triangulated graph is turned into a clique tree by finding the maximal cliques, where a clique is a set of nodes in which every pair of nodes is connected by an edge, and it is maximal if it is not properly contained within any other clique. In the clique tree, there is one node for each maximal clique of the triangulated graph and the edges connect nodes having variables in common. CTP works by passing messages around in the clique tree where the basic operation of the message passing from one clique to another is to sum out the variable that appears in one clique and does not appear in the other. The efficiency of the algorithm depends on the clique tree size, which strongly depends on the largest clique in the tree.

Variable elimination (VE) acts on a set of *factors*, functions which map each instantiation of variables to a non-negative number. The algorithm is based on the fact that variables can be summed out without having to construct the joint probability distribution explicitly. Thus, VE eliminates

the non-observed non-query variables one by one by summing them out. The *elimination order*, the ordering by which the variables are summed out, determines the number of numerical multiplications and numerical summations the algorithm performs, which determine the complexity of the algorithm.

The cost, in terms of the number of summations and multiplications, of answering a single query with no observations using CTP is of the same order of magnitude as using VE [128]. Therefore, in Chapter 2, to evaluate the efficiency of standard inference techniques in symmetric causal independence models when no observations have been received, we use only one of these algorithms. Due to its simplicity, we use the VE algorithm.

For complicated real-world problems, especially for problems involving a large number of variables and dependencies or variables taking values in a huge set of states, exact inference becomes intractable and approximation inference methods must be used. There has been much research devoted to finding efficient approximate inference algorithms, which resulted in many approximate inference algorithms relying on different key ideas. Two popular approaches include *sampling* and *variational* methods. The key idea of sampling techniques [80, 100] is that the joint distribution can be approximated by generating independent samples from it; whereas the basic idea of variational methods [56] is to formulate the computation of a marginal or conditional probability in terms of an optimization problem.

1.3 Parameter learning

Learning a Bayesian network from data includes two tasks: learning the structure and learning the parameters. If the structure is known or fixed, the problem of learning a Bayesian network reduces to learning the parameters. In this thesis, we deal with models whose structure is fixed; thus, in order to learn symmetric causal independence models, we only need to learn their parameters.

The task of learning parameters is not straightforward in situations where data is incomplete or the model has hidden variables. The common strategy to learn the parameters is to find the parameter set $\boldsymbol{\theta}$ that maximizes the *likelihood* that the observed data $\mathbf{D} = \{\mathbf{x}^1, \ldots, \mathbf{x}^N\}$ came from the

model:

$$L(\boldsymbol{\theta}) = \Pr(\mathbf{D} \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} \Pr(\mathbf{x}^i \mid \boldsymbol{\theta}).$$

A classical approach to find a maximum likelihood estimate for $\boldsymbol{\theta}$, the value of $\boldsymbol{\theta}$ which maximizes the likelihood, is the *expectation-maximization* (EM) algorithm [26]. The EM algorithm estimates the parameters by iteratively finding the expected values of the hidden variables and then finding the maximum likelihood estimate using the parameters from the expectation step. Each iteration of the algorithm increases the maximum likelihood estimate until a stable fixed point is reached.

Expectation-maximization is not one specific algorithm, it is a description of a class of related algorithms which make use of specific properties of the models. One of the first applications of the EM algorithm to learning the parameters in Bayesian networks was presented by Lauritzen [67].

1.4 Causal independence

The definition of a Bayesian network does not constrain how a variable depends on its parents. However, the number of conditional probabilities for a variable grows exponentially with the number of its parents, making the tasks of specifying the conditional probabilities and probabilistic inference in richly-connected Bayesian networks difficult or even intractable. Therefore, researchers proposed a number of ways of economically specifying the conditional probability of a variable with many parents.

Causal independence is a popular way to constrain the conditional probability tables for binary variables. The global structure of a causal independence model is shown in Figure 1.2; it expresses the idea that causes $C_1, \ldots, C_n$ influence a given common effect E through hidden variables $H_1, \ldots, H_n$ and a deterministic function f, called the *interaction function*. The hidden variable H_i is the contribution of cause variable C_i to the common effect E; most papers on causal independence models assume that absent causes do not contribute to this effect [43, 85]. The function f defines the way in which the hidden effects $H_1, \ldots, H_n$ - and, indirectly, the causes $C_1, \ldots, C_n$ - interact to yield the final effect E. The number of parameters in causal independence models is linear with the number of causes.

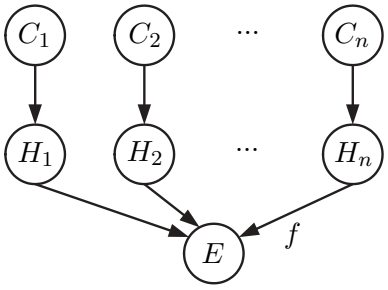


Figure 1.2: Causal independence model.

In practical applications of Bayesian networks, the function f is usually the logical OR function or the logical AND function. The resulting probabilistic models are called *noisy OR* and *noisy AND*; their underlying assumptions, respectively, are that the presence of at least one cause or the presence of all causes at the same time gives rise to the effect.

1.5 Carcinoid heart disease

In this thesis, we apply symmetric causal independence models to predict the development of carcinoid heart disease. In this section, we provide some information about this medical problem.

The body is composed of microscopic cells that are constantly degenerating, wearing out and being replaced by identical cells. Sometimes the cell replication process goes out of control, and a tumorous growth is formed. If the tumor is somewhat limited and does not spread to other areas or threaten to squeeze out or replace surrounding tissues, it is considered to be benign. However, if the growth is aggressive and threatens adjacent tissues or spreads to other locations in the body via lymph or blood, then the tumor is considered to be malignant. There are a few types of growths that are midway between the benign and malignant tumors. Carcinoid tumors, which start in the hormone-producing cells of various organs, are the most often occurring of these midway growths. These tumors have malignant potential but they usually grow slowly and most of them are asymptomatic through the natural lifetime and are discovered only upon surgery for unrelated reasons.

A small percent of carcinoid tumors, mostly faster-growing carcinoid tumors, secrete chemicals and hormones into the bloodstream causing the

carcinoid syndrome. In many cases, complications of carcinoid syndrome, namely, carcinoid crisis, bowel obstruction and carcinoid heart disease, are worse than the symptoms from the growth of the tumor. Carcinoid heart disease causes a thickening of the heart valves, making it difficult for them to function properly, which can eventually lead to heart failure. Carcinoid heart disease is the most dangerous complication of carcinoid syndrome as it occurs in over 65 percent of patients with carcinoid syndrome [79] and is a major source of morbidity and mortality for patients with carcinoid syndrome [19]. Given that so many carcinoid patients die of carcinoid heart disease, it is important to distinguish patients that are admitted to the clinic into patients who are likely to develop a severe form of carcinoid heart disease and those who are unlikely to develop this severe form. In this way, patients that are at risk can be given a more aggressive treatment in order to reduce the probability of the development of carcinoid heart disease.

1.6 Malaria

The other problem we try to learn more about by modelling it using symmetric causal independence models is gene regulation in malaria parasite.

Malaria is an infectious disease caused by a single-celled eukaryotic parasite, *Plasmodium*, which invades red blood cells and is transmitted by the bite of infected *Anopheles* mosquitoes. Malaria infects between 300 and 500 million people every year and accounts for more than one million deaths annually, with the vast majority of victims being young children in Africa [11, 93]. Survivors of severe malaria may be left with neurological effects including weakness in the limbs, speech disorders, behavioral disorders, blindness, hearing impairment and epilepsy. Death and secondary diseases are not the only impacts of malaria. The disease also creates a huge economic burden in malaria-endemic countries as the costs associated with malaria (expenditures on prevention, diagnosis, treatment and care of the disease as well as lost wages) are enormous [35]. Malaria also impedes economic development by limiting tourism, foreign investment and internal movement of labor and commerce.

Presently, there is no effective vaccine against malaria [109]. Methods used to control malaria include mosquito eradication and prophylactic drugs (most of which are also used for treatment of malaria). However, after prolonged exposure to an insecticide over several generations, it is com-

mon for mosquitoes to develop resistance, a capacity to survive contact with an insecticide [44]. Similarly, there is the alarming increase in *Plasmodium* resistance to commonly used anti-malarial drugs [38]. Because of the problems of resistance, new methods to control malaria are needed, and a vaccine holds the promise of controlling and perhaps eventually eradicating the disease.

Malaria has a complex life cycle, which consists of two main phases: a sexual phase in the mosquito and an asexual phase in the human host. Mosquitos become infected only if they take a blood meal from a person whose blood contains mature male and female stages of the parasite. During the sexual phase of the biological cycle of the parasite, the male and female gametocytes fuse in the mosquito's guts producing sporozoites, cells that infect new hosts, which migrate to mosquito's salivary glands. The asexual phase of the biological cycle of the parasite occurs as the mosquito injects her saliva into the human when it feeds. Within 30 minutes, the sporozoites are carried to the liver where they rapidly infect liver cells. Without causing symptoms, these sporozoites undergo a radical change and multiply furiously for the next 5-7 days. Tens of thousands of asexual stage merozoites are released from each infected liver cell, each of which rapidly target and invade a red blood cell. In the red blood cells, the parasites multiply to form new merozoites until the cells burst releasing large numbers of merozoites into the blood plasma, causing the characteristic fever associated with the disease. This phase of the disease occurs in cycles of approximately 48 hours. Some merozoites develop into gametocytes, the sexual stages of the parasite. If a mosquito bites this infected person and ingests the gametocytes, the malaria transmission cycle continues.

Basic knowledge of one of the processes fundamental to *Plasmodium* biology, i.e. gene regulation including transcriptional control, is still lacking. A thorough understanding of gene regulation in this organism is important for developing a better vaccine and identifying novel drug targets to fight this lethal disease.

1.7 Regulation of gene expression

The functions and properties of all cells are controlled by gene expression, the process by which the inheritable information in a gene is made into a functional gene product, mostly a protein (see Figure 1.3). Regulation of gene expression (also called gene regulation), which refers to the cellular

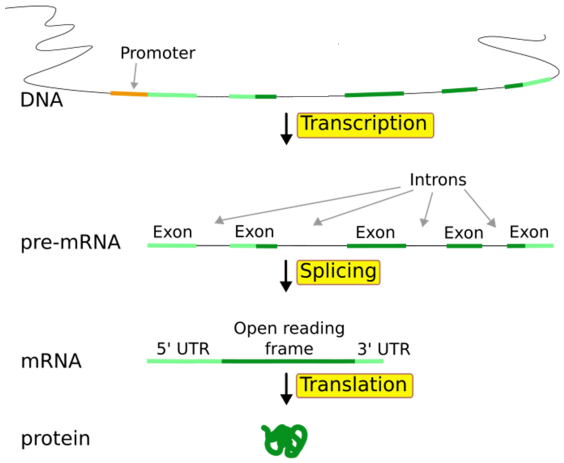


Figure 1.3: In the typical process of eukaryotic gene expression, a gene is first transcribed from DNA to the precursor mRNA (pre-mRNA). Then, the pre-mRNA undergoes major modifications, including the removal of introns, noncoding sequences, to become the messenger RNA (mRNA). The mRNA is sent out of the nucleus where the message is translated into proteins.

control of the amount and timing of changes to the appearance of the functional product of a gene, plays a vital role in single-celled microorganisms like *Plasmodium*, as it allows cells to adjust to their changing nutritional and physical environment.

Even though gene expression in eukaryotic cells is a chain of regulated events that together determine whether an active protein product is produced from a particular gene, gene expression is mostly controlled at the level of transcription - the first step in the production of proteins. In this way, cells can produce a particular mRNA only when the encoded protein is needed, thus minimizing wasted energy. Transcription of a gene is controlled by regulatory proteins - such as transcription factors - that bind to the regulatory sequences, short segments of DNA, which are usually positioned a short distance 'upstream' of the gene being regulated. Binding of activators to regulatory sequences called enhancers turns on transcription, and binding of repressors to other regulatory sequences called silencers turns off transcription. This protein-DNA interaction requires a binding site whose sequence pattern is more or less specific to each transcription factor. A concise representation, or a model, of the binding sites for a

transcription factor is referred to as a transcription factor binding motif.

1.8 Thesis outline

Many real-world Bayesian networks incorporate causal independence assumptions; however, only the noisy OR and noisy AND, two examples of causal independence models, are used in practice. Several authors proposed to expand the space of interaction functions in causal independence models by other symmetric Boolean functions. In this thesis, we study in detail causal independence models based on the symmetric Boolean functions, further referred to as symmetric causal independence models. The thesis is organized as follows.

In Chapter 2, we investigate the problem of probabilistic inference in symmetric causal independence models, with a special focus on causal independence models based on the Boolean threshold functions, further referred to as noisy threshold models. We establish a connection between the conditional probability distribution of the effect variable in these models and the Poisson binomial distribution. We investigate how the properties of the Poisson binomial distribution can be used for exact and approximate inference in symmetric causal independence models. We also compare the efficiency of the computational schemes developed with the efficiency of standard inference techniques.

The problem of learning the parameters in symmetric causal independence models is studied in Chapter 3. We present a computationally efficient EM algorithm to learn parameters in symmetric causal independence models, where the computational scheme of the Poisson binomial distribution is used to compute the conditional probabilities in the E-step. We study computational complexity and convergence of the developed algorithm. The presented EM algorithm allows us to assess the practical usefulness of symmetric causal independence models. The models are applied to a classification task and are shown to perform competitively with state-of-the-art classifiers.

Chapter 4 presents the application of the noisy threshold model to predict whether a patient with carcinoid syndrome will develop a carcinoid heart disease. We use data of fifty-four patients who suffered from a low-grade midgut carcinoid tumor, of which twenty-two patients developed carcinoid heart disease. The noisy threshold model performed favorably to other

state-of-the-art classification algorithms, and equally well as a decision-rule that was formulated by the physician.

In Chapter 5, we use noisy threshold models to identify regulatory sequence elements explaining membership to a gene expression cluster. Differently from other bioinformatics approaches, our method is able to model the logic behind gene regulation and to incorporate uncertainty about the functionality of putative regulatory sequence elements. We apply our approach to *Plasmodium falciparum*, the most virulant species of the four species of *Plasmodia* affecting humans. We have obtained several interesting results that deserve further (biological) investigation.

Chapter 6 concludes this thesis with a summary of main contributions and possible directions for further research.

Chapter 2

Probabilistic Inference

Causal independence modelling is a well-known method for reducing the size of probability tables, explaining the underlying mechanisms and enabling efficient inference in Bayesian networks. Many real-world Bayesian networks incorporate causal independence assumptions; however, only the noisy OR and noisy AND, two examples of causal independence models, are used in practice. Several authors proposed to expand the space of interaction functions in causal independence models by other symmetric Boolean functions. However, no research on the inference in the proposed models has yet been conducted. In this chapter, we investigate the problem of inference in causal independence models based on the Boolean threshold functions. We establish a connection between the conditional probability distribution of the effect variable in these models and the Poisson binomial distribution. We investigate how the properties of the Poisson binomial distribution can be used for computationally efficient exact and approximate inference in causal independence models based on the Boolean threshold functions. We also compare the efficiency of the computational schemes developed with the efficiency of standard inference techniques.

Parts of this chapter appeared in: R. Jurgelenaite, P.J.F. Lucas and T. Heskes. Exploring the noisy threshold function in designing Bayesian networks. In *Proceedings of AI-2005, the Twenty-fifth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence*, 2005.

2.1 Introduction

Bayesian networks [85] offer an appealing language for building models of domains with inherent uncertainty. However, because the number of conditional probabilities for each node grows exponentially with the number of its parents, it is usually unreliable if not infeasible to specify the conditional probabilities for nodes that have a large number of parents. Furthermore, for large, richly-connected Bayesian networks, probabilistic inference is difficult or even intractable. Causal independence modelling [27, 43, 105, 113, 126, 128] can greatly reduce the number of conditional probabilities to be assessed or elicited from experts and simplify the probabilistic inference. Another desirable property of causal independence models is their ability to explain the underlying relationships among the cause and effect variables.

Causal independence assumptions are often used in practical Bayesian network models [57, 103]. However, only the logical OR and AND operators are used in practice in defining the interaction among causes; their underlying assumption is that the presence of either at least one cause or all causes at the same time give rise to the effect. The resulting probabilistic submodels are called *noisy OR* and *noisy AND*, respectively. Some authors proposed to expand the space of possible interaction functions in causal independence models by considering other symmetric Boolean functions: the idea was mentioned but not developed further in [76]; an analysis of the qualitative patterns of these new models was presented in [71]. The generalization preserves efficiency and understandability of the noisy-OR and noisy-AND models, while at the same time allowing more flexibility in modelling the interaction among causes.

In this chapter, we explore the inference problem in causal independence models with a symmetric Boolean function. It is known that any symmetric Boolean function can be decomposed into threshold functions, i.e. Boolean functions that return truth if the number of their true arguments is greater than or equal to a given positive integer [120]. Thus, threshold functions offer a natural basis for the analysis of causal independence models. Likewise, the study of probabilistic inference in causal independence models with threshold functions acts as a basis for the analysis of inference in any causal independence models based on symmetric Boolean functions. Inference in causal independence models with the threshold interaction function is the main topic of this chapter. These models will further be referred to as *noisy threshold models*. We establish a connection

between conditional probability distributions in the noisy threshold model and Poisson binomial distribution. This connection enables us to explore the properties of noisy threshold models given the properties of this well-studied probability distribution. We also investigate how the standard inference techniques can be applied for the probabilistic inference in the noisy threshold models.

The structure of this chapter is as follows. In the following section, Bayesian networks, causal independence models and Boolean functions are reviewed. Section 2.3 presents a basis for inference in the noisy threshold models. In Section 2.4, we establish a connection between the conditional probability distribution of the effect variable in the noisy threshold models and the Poisson binomial distribution. In section 2.5, we study the exact methods to compute the conditional probability distribution of the effect in the noisy threshold models, while Section 2.6 presents and investigates approximation and bounding techniques. Finally, in Section 2.7, we summarize what has been achieved by this research.

2.2 Preliminaries

2.2.1 Bayesian networks

A *Bayesian network* $\mathcal{B} = (G, \Pr)$ represents a factorized joint probability distribution on a set of random variables $\mathbf{V}$. It consists of two parts: (1) a qualitative part, represented as an acyclic directed graph (ADG) $G = (\mathbf{V}(G), \mathbf{A}(G))$, where there is a one-to-one correspondence between the vertices $\mathbf{V}(G)$ and the random variables in $\mathbf{V}$, and the arcs $\mathbf{A}(G)$ represent the conditional dependencies between the variables; (2) a quantitative part Pr consisting of local probability distributions $\Pr(V \mid \pi(V))$, for each variable $V \in \mathbf{V}$ given the parents $\pi(V)$ of the corresponding vertex (interpreted as variables). The joint probability distribution Pr is factorized according to the structure of the graph as

$$\Pr(\mathbf{V}) = \prod_{V \in \mathbf{V}} \Pr(V \mid \pi(V)).$$

Each variable $V \in \mathbf{V}$ has a finite set of mutually exclusive states. In this chapter, we assume all variables to be binary; we will use v to denote the

realization of the random variable V, v^+ to denote $V = \top$ (true) and v^- to denote $V = \bot$ (false).

2.2.2 Causal modelling and Boolean functions

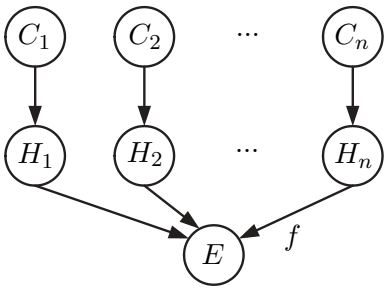


Figure 2.1: Causal independence model.

Causal independence (also known as independence of causal influence) is a popular way to specify interactions among cause variables. The global structure of a causal independence model is shown in Figure 2.1; it expresses the idea that causes $C_1, \ldots, C_n$ influence a given common effect E through hidden variables $H_1, \ldots, H_n$ and a deterministic function f, called the *interaction function*. The hidden variable H_i is considered to be the contribution of the cause variable C_i to the common effect E. The function f represents the way in which the hidden effects H_i and indirectly also the causes C_i interact to yield the final effect E. The function f is defined in such a way that when a relationship, as modelled by the function f, between $H_i, i = 1, \ldots, n$, and $E = \top$ is satisfied, then it holds that $f(h_1, \ldots, h_n) = \top$. It is assumed that $\Pr(e^+ \mid h_1, \ldots, h_n) = 1$ if $f(h_1, \ldots, h_n) = \top$, and $\Pr(e^+ \mid h_1, \ldots, h_n) = 0$ if $f(h_1, \ldots, h_n) = \bot$.

A causal independence model is defined in terms of the causal parameters $\Pr(H_i \mid C_i), i = 1, \ldots, n$ and the function $f(h_1, \ldots, h_n)$. Most papers on causal independence models assume that absent causes do not contribute to the effect [43, 85]. In terms of probability theory, this implies that it holds that $\Pr(h_i^+ \mid c_i^-) = 0$; as a consequence, it holds that $\Pr(h_i^- \mid c_i^-) = 1$. We make the same assumption in this chapter.

In situations in which a model does not capture all possible causes, it is useful to introduce a *leaky cause* which captures the unidentified causes contributing to the effect and is assumed to be always present [45]. In an

arithmetic context, the leaky cause is handled in the same way as other causes. In Chapter 3, we model the leak term by adding an additional input $C_{n+1} = 1$ to the data; in an arithmetic context the leaky cause is treated in the same way as identified causes.

The conditional probability of the effect E given the causes $C_1, \ldots, C_n$ is obtained from the causal parameters $\Pr(H_i \mid C_i)$ [71, 128]:

$$\Pr(e \mid c_1, \ldots, c_n) = \sum_{f(h_1,\ldots,h_n)=e} \prod_{i=1}^{n} \Pr(h_i \mid c_i). \tag{2.1}$$

In this thesis, we assume that the function f in Equation (2.1) is a Boolean function. Systematic analyses of the global probabilistic patterns in causal independence models based on some of the Boolean functions were presented in [71]. However, there are 2^{2^n} different n-ary Boolean functions [31, 120]; thus, the potential number of causal interaction models is huge. However, if we assume that the order of the cause variables does not matter, the Boolean functions become *symmetric* [120] and the number of functions reduces to 2^{n+1}.

An important symmetric Boolean function is the *exact* Boolean function ϵ_l, which has function value true, i.e. $\epsilon_l(h_1, \ldots, h_n) = \top$, if $\sum_{i=1}^{n} \nu(h_i) = l$ with $\nu(h_i)$ equal to 1 if h_i is equal to true and 0 otherwise. A symmetric Boolean function can be decomposed in terms of the exact functions ϵ_l as follows [120]:

$$f(h_1, \ldots, h_n) = \bigvee_{i=0}^{n} \epsilon_i(h_1, \ldots, h_n) \wedge \gamma_i, \tag{2.2}$$

where γ_i are Boolean constants that depend on the function f. For example, for the Boolean function defined in terms of the OR operator, we have: $\gamma_0 = \bot$ and $\gamma_1 = \ldots = \gamma_n = \top$.

Another useful symmetric Boolean function is the *threshold* function τ_k, which simply checks whether there are at least k trues among the arguments, i.e. $\tau_k(h_1, \ldots, h_n) = \top$, if $\sum_{i=1}^{n} \nu(h_i) \geq k$ with $\nu(h_i)$ equal to 1 if h_i is equal to true and 0 otherwise. To express it in the Boolean constants, we have: $\gamma_0 = \cdots = \gamma_{k-1} = \bot$ and $\gamma_k = \cdots = \gamma_n = \top$. Note that the OR function corresponds to the threshold function τ_1, and the AND function corresponds to the threshold function τ_n. Hence, these two commonly used Boolean functions are the extremes of a spectrum of the Boolean threshold functions. Any exact Boolean function can be written as the subtraction

of two threshold functions; therefore, any symmetric Boolean function can be decomposed into threshold functions. To simplify the further discussion, causal independence models based on the Boolean threshold function and the symmetric Boolean function will be referred to as *noisy threshold models* and *symmetric causal independence models*, respectively.

Figure 2.2 shows a noisy threshold model that represents a real-world medical problem of the response to the treatment of patients with gastric non-Hodgkin's lymphoma. Gastric non-Hodgkin lymphoma is a type of cancer of the lymphatic system, the disease-fighting network spread throughout the body, which originates in the stomach. The early outcome of the treatment, which is the effect in the model, denotes the endoscopically verified result of the treatment, six to eight weeks after treatment; the positive state of this variable, complete remission, defines a situation in which all clinical signs of disease disappear with the treatment. The following pretreatment prognostic factors are available: (1) age; (2) general health status; (3) bulky disease; (4) histological classification; (5) stage of the cancer; (6) clinical signs (hemorrhage, perforation or obstruction due to the disease); (7) leaky cause that summarizes the prognostic factors that are not included into the model. The prognostic factors correspond to the cause variables in the model. A more elaborate description of the model will be presented in Chapter 3.

In the following, we explore the inference problem in the noisy threshold models.

2.3 Inference in the noisy threshold models

Inference refers to the process of computing conditional and marginal probabilities in graphical models. The inference problem in causal independence models can be subdivided into two groups of queries: (1) computing probabilities of the effect variable, and (2) computing probabilities of a subset of the cause variables.

Let Q and E be disjoint subsets of the cause variable indices of a causal independence model such that $\mathbf{C}_Q$ and $\mathbf{C}_E$ are disjoint subsets of the query and evidence cause variables, respectively. Then, the conditional probability of the effect variable and the query cause variables can be

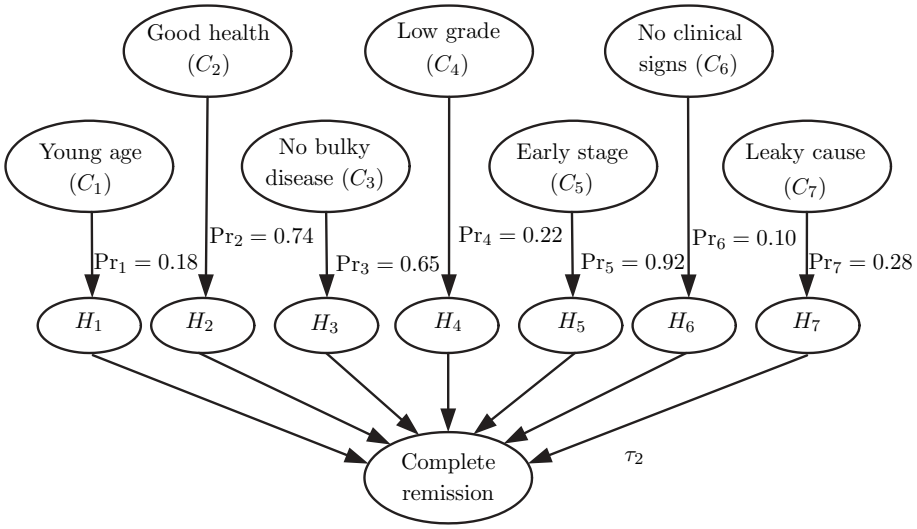


Figure 2.2: Noisy threshold model modelling the complete remission following the treatment of gastric Non-Hodgkin lymphoma. $\Pr_k$ is a shorthand for $\Pr(h_k^+ \mid c_k^+)$, and τ_2 denotes a Boolean threshold function with $k = 2$. The leaky cause C_7 is assumed to be always active.

written as

$$\Pr(e, \mathbf{c}_Q \mid \mathbf{c}_E) = \Pr(e \mid \mathbf{c}_Q, \mathbf{c}_E)\Pr(\mathbf{c}_Q \mid \mathbf{c}_E). \tag{2.3}$$

Given the assumption that the cause variables of a causal independence model are independent when the effect is not observed, Equation (2.3) becomes

$$\Pr(e, \mathbf{c}_Q \mid \mathbf{c}_E) = \Pr(e \mid \mathbf{c}_Q, \mathbf{c}_E) \prod_{i \in Q} \Pr(c_i). \tag{2.4}$$

Using Bayes' rule, we can write the conditional probability of the query cause variables of a causal independence model as

$$\Pr(\mathbf{c}_Q \mid e, \mathbf{c}_E) = \frac{\Pr(e \mid \mathbf{c}_Q, \mathbf{c}_E)\Pr(\mathbf{c}_Q \mid \mathbf{c}_E)}{\Pr(e \mid \mathbf{c}_E)}. \tag{2.5}$$

Following the assumption of independence of causes when the effect is not observed, we obtain

$$\Pr(\mathbf{c}_Q \mid e, \mathbf{c}_E) = \frac{\Pr(e \mid \mathbf{c}_Q, \mathbf{c}_E) \prod_{i \in Q} \Pr(c_i)}{\Pr(e \mid \mathbf{c}_E)}. \tag{2.6}$$

As we can see from Equations (2.4) and (2.6), inference in causal independence models boils down to the computation of the probability $\Pr(e \mid \mathbf{c}_Q, \mathbf{c}_E)$. Note that computing the conditional probability of the query cause variables $\mathbf{c}_Q$ requires computing it for every realization of $\mathbf{c}_Q$, which amounts to instantiating the query causes. Therefore, from now on, we view the query causes as evidence causes and investigate the conditional probability $\Pr_{\tau_k}(e \mid \mathbf{c}_E)$. To compute this conditional probability, the cause variables indexed by $R = \{1, \ldots, n\} \setminus E$ are variables that have to be marginalized out. Using the assumption of independence of causes, we have

$$\begin{aligned}
\Pr(e \mid \mathbf{c}_E) &= \sum_{\mathbf{c}_R} \Pr(e, \mathbf{c}_R \mid \mathbf{c}_E) \\
&= \sum_{\mathbf{c}_R} \Pr(\mathbf{c}_R) \sum_{f(h_1,\ldots,h_n)=e} \prod_{j \in E} \Pr(h_j \mid c_j) \prod_{l \in R} \Pr(h_l \mid c_l) \\
&= \sum_{f(h_1,\ldots,h_n)=e} \prod_{j \in E} \Pr(h_j \mid c_j) \prod_{l \in R} \sum_{c_l} \Pr(h_l \mid c_l) \Pr(c_l) \\
&= \sum_{f(h_1,\ldots,h_n)=e} \prod_{j \in E} \Pr(h_j \mid c_j) \prod_{l \in R} \Pr(h_l). \qquad (2.7)
\end{aligned}$$

Combining (2.2) and (2.7), we obtain

$$\Pr(e^+ \mid \mathbf{c}_E) = \sum_{\substack{0 \le i \le n \\ \gamma_i}} \sum_{\epsilon_i(h_1,\ldots,h_n)=\top} \prod_{j \in E} \Pr(h_j \mid c_j) \prod_{l \in R} \Pr(h_l). \tag{2.8}$$

Hence, Equation (2.8) yields a general formula to compute the probability of the effect in terms of the Boolean exact functions in any symmetric causal independence model.

Let us denote a conditional probability of the effect given the evidence cause variables in a noisy threshold model with interaction function τ_k as

$\Pr_{\tau_k}(e \mid \mathbf{c}_Q, \mathbf{c}_E)$. Then, from Equation (2.8), it follows that

$$\Pr_{\tau_k}(e^+ \mid \mathbf{c}_E) = \sum_{k \le i \le n} \sum_{\epsilon_i(h_1,\ldots,h_n)=\top} \prod_{j \in E} \Pr(h_j \mid c_j) \prod_{l \in R} \Pr(h_l). \quad (2.9)$$

Since the effect variable is binary, we will further consider only the computation of the probability $\Pr_{\tau_k}(e^+ \mid \mathbf{c}_E)$.

2.4 The Poisson binomial distribution and noisy threshold models

It turns out that the conditional probability distribution of the effect variable in the noisy threshold models is closely connected to the Poisson binomial distribution.

Let l denote the number of successes in n independent trials, where p_i is a probability of success in the i-th trial, $i = 1, \ldots, n$; let $\mathbf{p} = (p_1, \ldots, p_n)$. The trials are then called *Poisson trials* [32], and $\mathrm{B}(l; \mathbf{p})$ denotes the *Poisson binomial distribution* (also known as the distribution of the number of successes of independent trials) [30, 69]:

$$\mathrm{B}(l; \mathbf{p}) = \left\{ \prod_{i=1}^{n} (1 - p_i) \right\} \sum_{1 \le j_1 < \ldots < j_l \le n} \prod_{z=1}^{l} \frac{p_{j_z}}{1 - p_{j_z}}. \quad (2.10)$$

The Poisson trials are characterized by the mean $\mu = \frac{1}{n} \sum_{i=1}^{n} p_i$ and the variance $\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (p_i - \mu)^2$. When the variance $\sigma^2 = 0$, i.e. the success probability p_i is a constant p, the trials are called Bernoulli trials and $\mathrm{B}(l; \mathbf{p})$ reduces to the binomial distribution: $\mathrm{B}(l; p) = \binom{n}{l} p^l (1 - p)^{n-l}$.

Let us define a vector of probabilistic parameters $\mathbf{p}(\mathbf{c}_E) = (p_1, \ldots, p_n)$ with

$$p_i = \begin{cases} \Pr(h_i^+ \mid c_i) & \text{if } i \in E, \\ \Pr(h_i^+) & \text{otherwise.} \end{cases}$$

The connection between the Poisson binomial probabilities and the conditional probabilities of the effect with the Boolean exact function as an interaction function is as stated in the following proposition.

Proposition 1 *It holds that*

$$\sum_{\epsilon_i(h_1,\ldots,h_n)=\top} \prod_{j\in E} \Pr(h_j \mid c_j) \prod_{l\in R} \Pr(h_l) = \mathrm{B}(i; \mathbf{p}(\mathbf{c}_E)). \tag{2.11}$$

Proof: Note that in Equation (2.8), the sum

$$\sum_{\epsilon_i(h_1,\ldots,h_n)=\top} \prod_{j\in E} \Pr(h_j \mid c_j) \prod_{l\in R} \Pr(h_l)$$

was defined as the probability that exactly i hidden variables $H_1, \ldots, H_n$ are true. A hidden variable H_m can be seen as an independent trial which has a probability of success that equals $\Pr(h_m^+ \mid c_m)$ when C_m is fixed at some value and $\Pr(h_m^+)$ otherwise. Combining the definition of the vector $\mathbf{p}(\mathbf{c}_E)$ and the definition of the Poisson binomial distribution, the result in the premise of this proposition is obtained. ■

Now, we can establish the connection between the conditional probability of the effect in the noisy threshold model and the Poisson binomial distribution.

Proposition 2 *It holds that*

$$\Pr_{\tau_k}(e^+ \mid \mathbf{c}_E) = \sum_{k\leq i\leq n} \mathrm{B}(i; \mathbf{p}(\mathbf{c}_E)). \tag{2.12}$$

Proof: The proof follows from Equation (2.8) and Proposition 1. ■

Let $\rho(\mathbf{c}_E)$ denote the number of probabilistic parameters $p \in \mathbf{p}(\mathbf{c}_E), p \neq 0$. If this number of possibly 'active' hidden variables $\rho(\mathbf{c}_E)$ is smaller than the threshold k, the conditional probability of the effect equals zero as is shown in the following corollary.

Corollary 3 *Let $\rho(\mathbf{c}_E) < k, 1 \leq k \leq n$, then $\Pr_{\tau_k}(e^+ \mid \mathbf{c}_E) = 0$.*

Proof: From the definition of the Poisson binomial distribution it follows that $\mathrm{B}(l; \mathbf{p}(\mathbf{c}_E)) = 0$ for all $l > \rho(\mathbf{c}_E)$. The required result, therefore, follows directly from Proposition 2. ■

In the next two sections, we will review the exact, approximation and bounding methods to compute the conditional probability $\Pr_{\tau_k}(e^+ \mid \mathbf{c}_E)$.

2.5 Exact methods to compute the probability $\Pr_{\tau_k}(e^+ \,|\, \mathbf{c}_E)$

This section investigates how standard inference techniques and recursive methods to compute the Poisson binomial distribution can be used to calculate the probability $\Pr_{\tau_k}(e^+ \mid \mathbf{c}_E)$. We also assess the computational cost of these methods in terms of number of multiplications and additions. We consider the most expensive case in terms of computational cost, we assume that $E = \emptyset$.

2.5.1 Standard inference techniques

The most common exact inference methods are clique tree propagation [54, 68, 101], which performs belief propagation on a modified graph called a junction tree, and variable elimination [127], which eliminates the non-observed non-query variables one by one summing them out. To perform exact inference in causal independence models using one of these standard approaches, the interaction function has to be converted into a conditional distribution. A trivial conversion for a symmetric Boolean function is

$$\Pr(e^+ \mid h_1, \ldots, h_n) = \begin{cases} 1 & \text{if } \gamma_l = \top \text{ where } l = \sum_{i=1}^{n} \nu(h_i), \\ 0 & \text{otherwise.} \end{cases}$$

However, both above mentioned inference methods have a computational complexity that is exponential in the so-called treewidth, which is one less than the cardinality of the largest elimination clique. The triangulated graph of the Bayesian network obtained using trivial conversion contains a clique with cardinality $n + 1$. Luckily, it is possible to decompose the interaction function in order to obtain a smaller treewidth. The two well known approaches to the decomposition, *parent divorcing* [83] and *temporal transformation* [42], construct Bayesian networks such that the cardinality of their maximal cliques equals 3. The Bayesian networks constructed by these two approaches are shown in Figures 2.3 and 2.4.

Following the same kind of reasoning as in [128], we investigate the efficiency of inference in symmetric causal independence models using standard inference techniques. To do so, we convert the models to Bayesian networks using parent divorcing and temporal transformation and, then,

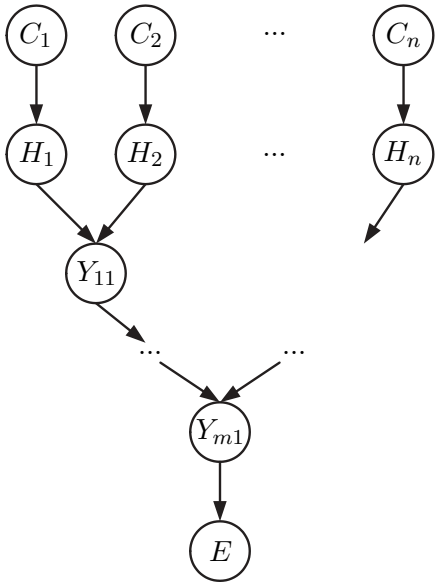


Figure 2.3: Parent divorcing transformation of a symmetric causal independence model.

use variable elimination for inference. Table 2.1 shows the number of multiplications and additions to compute the probability $\Pr(e^+)$ in a symmetric causal independence model. We consider two cases: in the general case, we do not use knowledge of the properties of causal independence models; in the adjusted case, we take these properties into account. The specific properties of symmetric causal independence models include the assumption that an absent cause does not contribute to the effect and the structure of the conditional probability table $\Pr(Y_i \mid Y_{i-1}, H_i)$, which contains many zeros. To evaluate the efficiency of parent divorcing, we derived formulas for the models with a number of cause variables $n = 2^m, m \in \mathbb{Z}^+$. This corresponds to the 'best case' for the parent divorcing approach, as the computational complexity for n which is not a power of two is slightly higher. The formulas presented in Table 2.1 show that parent divorcing is somewhat more efficient than the temporal transformation. However, the differences are pretty small, the methods differ from each other by at most a factor of 2.

When a symmetric Boolean function is the threshold function, the number of probabilities in Bayesian networks obtained using the two transformations can be reduced. Firstly, the number of states for the variables

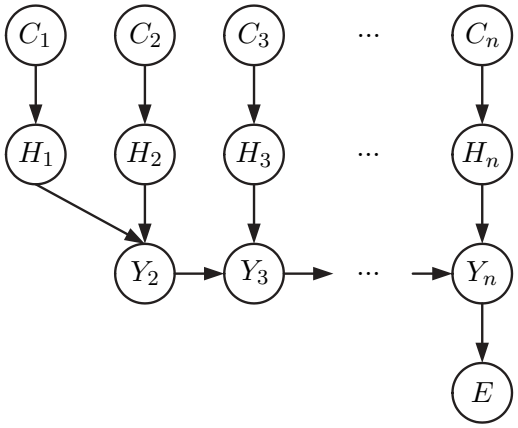


Figure 2.4: Temporal transformation of a symmetric causal independence model.

$Y_2, \ldots, Y_n$ or $Y_{1,1}, \ldots, Y_{m,1}$ can be reduced by introducing the states $\top$ and $\bot$ which represent k or more hidden variables among ancestors that are equal to true and $n-k+1$ or more hidden variables among ancestors that are equal to false, respectively. Secondly, the variable E can be removed, as the variable Y_n has only two states, $\top$ and $\bot$, and effectively models the effect in this updated model. Table 2.2 shows the cost of computing the probability $\Pr(e^+)$ in the noisy threshold model; for the parent divorcing approach, the formulas are derived assuming that the threshold k is a power of 2. Since the expressions presented in Table 2.2 are rather complicated, we computed the number of required operations to calculate $\Pr_{\tau_k}(e^+)$ for specific values of k. The leading terms of the derived expressions are given in Table 2.3. The differences in efficiency between the parent divorcing and the temporal transformation approaches are even smaller than in causal independence models with any symmetric Boolean function.

An alternative way to perform inference in the noisy threshold model is to represent it as a combination of a noisy adder model and a threshold function. *Heterogeneous factorization* of the resulting model enables efficient inference using the VE_1 algorithm [128]. The VE_1 algorithm combines the heterogeneous factors pairwise, eliminating one at a time. Therefore, given the same knowledge about the model, the VE_1 algorithm requires the same number of multiplications and additions to compute the probability $\Pr(e^+)$ as the adjusted temporal transformation method.

Table 2.1: The cost of computing the probability $\Pr(e^+)$ in a symmetric causal independence model decomposed using parent divorcing (PD) and temporal transformation (TT).

Method	Multiplications	Additions
PD (general)	$\frac{2}{3}n^3 + 5n^2 + (\frac{4}{3} + 4\log_2 n)n - 1$	$\frac{1}{3}n^3 + \frac{5}{2}n^2 + (\frac{1}{6} + \log_2 n)n$
TT (general)	$\frac{4}{3}n^3 + 4n^2 + \frac{23}{3}n - 7$	$\frac{2}{3}n^3 + \frac{3}{2}n^2 + \frac{17}{6}n - 2$
PD (adjusted)	$\frac{1}{2}n^2 + (\frac{5}{2} + \log_2 n)n$	$\frac{1}{2}n^2 + \frac{3}{2}n$
TT (adjusted)	$n^2 + 3n - 1$	$\frac{1}{2}n^2 + \frac{3}{2}n$

Table 2.2: The cost of computing the probability $\Pr_{\tau_k}(e^+)$ where $k \leq \frac{n}{2}$. The formulas for the probability $\Pr_{\tau_k}(e^+)$ with $k > \frac{n}{2}$ can be obtained by replacing k with $n - k + 1$.

Method	Multiplications	Additions
PD (general)	$\left(\frac{8}{3}(k+1)^2 + \frac{17}{3}k + 4\log_2 k + \frac{11}{3}\right) n$ $-4k(k+1)^2$	$\left(\frac{4}{3}(k+1)^2 + \frac{17}{6}k + \log_2 k - \frac{1}{6}\right) n$ $-2k^2(k+2)$
TT (general)	$4(k^2+2k+2)n$ $-4\left(\frac{4}{3}k^3 + k^2 - \frac{4}{3}k + 3\right)$	$(2k^2+3k+3)n - \left(\frac{8}{3}k^3 + k^2 - \frac{5}{3}k + 4\right)$
PD (adjusted)	$\left(\frac{3}{2}k + \log_2 k + \frac{3}{2}\right) n - (k^2+1)$	$(\frac{3}{2}k - \frac{1}{2} + \frac{2}{k})n - (k^2 - 2k + 3)$
TT (adjusted)	$(2k+1)n - 2(k^2 - k + 1)$	$(k+1)n - (k^2 - k + 1)$

Table 2.3: The leading terms of the cost of computing the probability $\Pr_{\tau_k}(e^+)$ for specific values of k.

Method	Multiplications			Additions		
	$k=1$	$k=\frac{n}{4}$	$k=\frac{n}{2}$	$k=1$	$k=\frac{n}{4}$	$k=\frac{n}{2}$
PD (general)	$20n$	$\frac{5}{48}n^3$	$\frac{1}{6}n^3$	$8n$	$\frac{5}{96}n^3$	$\frac{1}{12}n^3$
TT (general)	$20n$	$\frac{1}{6}n^3$	$\frac{1}{3}n^3$	$8n$	$\frac{1}{12}n^3$	$\frac{1}{6}n^3$
PD (adjusted)	$3n$	$\frac{5}{16}n^2$	$\frac{1}{2}n^2$	$3n$	$\frac{5}{16}n^2$	$\frac{1}{2}n^2$
TT (adjusted)	$3n$	$\frac{3}{8}n^2$	$\frac{1}{2}n^2$	$2n$	$\frac{3}{16}n^2$	$\frac{1}{4}n^2$

2.5.2 Recursive methods to compute the Poisson binomial distribution

So far, we have considered the use of standard Bayesian networks inference techniques. An alternative is to use the properties of the Poisson binomial distribution.

Computing the probability $\mathrm{B}(l;\mathbf{p})$ naively, e.g. through (3.2), one needs to sum $\frac{n!}{l!(n-l)!}$ terms, which is impractical even when l and n are of moderate sizes. Recursive formulas require a smaller number of operations for computing this sum. At least two recursive methods to compute the Poisson binomial probabilities have been reported.

The first recursive method to calculate the Poisson binomial distribution was presented by Howard [49]. The Poisson binomial probability $\mathrm{B}(l;\mathbf{p})$ is computed recursively by

$$\mathrm{B}(l;(p_1,\ldots,p_n)) = \mathrm{B}(l;(p_1,\ldots,p_{n-1}))(1-p_n) + \mathrm{B}(l-1;(p_1,\ldots,p_{n-1}))p_n.$$

The computational cost of this method is identical to the cost of the adjusted temporal transformation and the heterogeneous factorization of the noisy adder model with a threshold function.

The second recursive method to compute the Poisson binomial distribution was presented by Chen et al. [15]. This method is slightly less efficient than the previously discussed methods; therefore, we do not explain it here.

2.5.3 Noisy threshold models for classification

Note that all methods for exact inference in the noisy threshold models are quadratic in the number of cause variables n. However, if the effect variable is the only query variable, a method linear in n often suffices as we will show.

Classification is one of the possible applications of the noisy threshold models. The causes can be interpreted as feature variables, the effect as the class variable, and the conditional probability $\Pr_{\tau_k}(e \mid \mathbf{c}_E)$ as the class probability. For binary classifiers, the default *classification threshold* (not to be confused with the threshold function) is typically set to $\frac{1}{2}$. To classify a data instance using a noisy threshold classifier as defined, we do not need to compute the exact probability $\Pr_{\tau_k}(e^+ \mid \mathbf{c}_E)$, it is enough to know whether $\Pr_{\tau_k}(e^+ \mid \mathbf{c}_E) \leq \frac{1}{2}$ or $\Pr_{\tau_k}(e^+ \mid \mathbf{c}_E) \geq \frac{1}{2}$. We will show that in most cases there is a simple way to determine which state of the effect/class variable is more likely to occur.

We start by introducing some properties of the Poisson binomial distribution that are needed to derive this result.

The *mean* $m(\mathbf{p})$ of the distribution $\mathrm{B}(i; \mathbf{p})$ is by definition equal to

$$m(\mathbf{p}) = \sum_{i=0}^{n} i \, \mathrm{B}(i; \mathbf{p}).$$

By means of some algebraic manipulation, it can be shown that the mean $m(\mathbf{p})$ of the Poisson binomial distribution $\mathrm{B}(l; \mathbf{p})$ is equal to the sum of the probabilities of success [32]

$$m(\mathbf{p}) = \sum_{i=1}^{n} p_i.$$

The *median* $M(\mathbf{p})$ of the distribution $\mathrm{B}(l; \mathbf{p})$ is the integer number such that

$$(1) \quad \sum_{i=0}^{M(\mathbf{p})} \mathrm{B}(i; \mathbf{p}) \geq \frac{1}{2},$$

$$(2) \quad \sum_{i=M(\mathbf{p})}^{n} \mathrm{B}(i;\mathbf{p}) \geq \frac{1}{2}.$$

Jogdeo and Samuels [55] established a connection between the mean $m(\mathbf{p})$ and the median $M(\mathbf{p})$ of the Poisson binomial distribution:

$$M(\mathbf{p}) = \begin{cases} l & \text{if } m(\mathbf{p}) = l \\ l \text{ or } l+1 & \text{if } l < m(\mathbf{p}) < l+1 \end{cases} \tag{2.13}$$

where $0 \leq l \leq n$ is an integer.

Knowing this connection between the median and the mean, we can distinguish between the conditional probabilities where the effect E is more likely to be present and the conditional probabilities where the effect E is more likely to be absent.

Proposition 4 *Let* $\mathbf{m}(\mathbf{p}(\mathbf{c}_E))$ *and* $M(\mathbf{p}(\mathbf{c}_E))$ *be the mean and the median of the probability distribution* $\mathrm{B}(i;\mathbf{p}(\mathbf{c}_E))$*, respectively. Let* $\rho(\mathbf{c}_E) \geq k, 1 \leq k \leq n$*, then*

- $\Pr_{\tau_k}(e^+ \mid \mathbf{c}_E) \geq \frac{1}{2}$ *if* $m(\mathbf{p}(\mathbf{c}_E)) \geq k$*,*
- $\Pr_{\tau_k}(e^+ \mid \mathbf{c}_E) \leq \frac{1}{2}$ *if* $m(\mathbf{p}(\mathbf{c}_E)) \leq k-1$*.*

Proof: Equation (2.12) can be written in the form

$$\begin{aligned} \Pr_{\tau_k}(e^+ \mid \mathbf{c}_E) &= \sum_{i=k}^{M(\mathbf{p}(\mathbf{c}_E))-1} \mathrm{B}(i;\mathbf{p}(\mathbf{c}_E)) + \sum_{i=M(\mathbf{p}(\mathbf{c}_E))}^{\rho(\mathbf{c}_E)} \mathrm{B}(i;\mathbf{p}(\mathbf{c}_E)) \\ &\quad \text{if } M(\mathbf{p}(\mathbf{c}_E)) \geq k, \\ \Pr_{\tau_k}(e^+ \mid \mathbf{c}_E) &= 1 - \sum_{i=0}^{M(\mathbf{p}(\mathbf{c}_E))} \mathrm{B}(i;\mathbf{p}(\mathbf{c}_E)) - \sum_{i=M(\mathbf{p}(\mathbf{c}_E))+1}^{k-1} \mathrm{B}(i;\mathbf{p}(\mathbf{c}_E)) \\ &\quad \text{if } M(\mathbf{p}(\mathbf{c}_E)) \leq k-1. \end{aligned}$$

Then from the definition of the median $M(\mathbf{p}(\mathbf{c}_E))$ *we get the following inequalities:*

$$\begin{aligned} &\Pr_{\tau_k}(e^+ \mid \mathbf{c}_E) \geq \frac{1}{2} && \text{if } M(\mathbf{p}(\mathbf{c}_E)) \geq k, \\ &\Pr_{\tau_k}(e^+ \mid \mathbf{c}_E) \leq \frac{1}{2} && \text{if } M(\mathbf{p}(\mathbf{c}_E)) \leq k-1 \end{aligned}$$

From Equation (2.13) it follows that $M(\mathbf{p}(\mathbf{c}_E))$ *equals* $\lfloor m(\mathbf{p}(\mathbf{c}_E)) \rfloor$ *or* $\lceil m(\mathbf{p}(\mathbf{c}_E)) \rceil$*, hence the inequalities above can be written as*

$$\begin{aligned} \Pr_{\tau_k}(e^+ \mid \mathbf{c}_E) &\geq \frac{1}{2} \qquad \text{if } m(\mathbf{p}(\mathbf{c}_E)) \geq k, \\ \Pr_{\tau_k}(e^+ \mid \mathbf{c}_E) &\leq \frac{1}{2} \qquad \text{if } m(\mathbf{p}(\mathbf{c}_E)) \leq k-1. \end{aligned}$$

■

Let us return to the model of gastric non-Hodgkin lymphoma from Section 2.2.2. Assuming there is no missing data, i.e. $E = \{1, \ldots, 6\}$, there are $2^6 = 64$ possible realizations of $\mathbf{c}_E$. Applying Proposition 4, we obtain that 23 of the resulting conditional probabilities $\Pr(e^+ \mid \mathbf{c}_E)$ are equal or larger than $\frac{1}{2}$ and 9 probabilities are equal or smaller than $\frac{1}{2}$. Unfortunately, the other 32 probabilities cannot be determined as their corresponding $m(\mathbf{p}(\mathbf{c}_E))$ value falls into the interval $(1, 2)$. However, this result strongly depends on the size of the model, i.e. in a larger noisy threshold model where $m(\mathbf{p}(\mathbf{c}_E))$ values vary more, a larger percentage of the conditional probabilities of the effect can be classified based solely on the mean $m(\mathbf{p}(\mathbf{c}_E))$.

2.6 Approximate methods to compute the probability $\Pr_{\tau_k}(e^+ \mid \mathbf{c}_E)$

In Section 2.5, we presented a number of exact methods to compute the probability $\Pr_{\tau_k}(e^+ \mid \mathbf{c}_E)$ in a quadratic number of operations with respect to n. However, for some practical applications, the number of cause variables n is in the hundreds. An example of such a real-wold application is the noisy threshold model for classification of the documents from the Reuters data collection, see Chapter 3. Some document classes in this data collection have hundreds of relevant features, which are modelled as cause variables. Another example of such a practical application is the QMR-DT model [103]. Although the interaction among findings and diseases is modelled by means of logical OR, simulation algorithms are used to perform inference in this model. Even for problems with a smaller number of causes, linear, rather than quadratic, complexity can be the difference between feasible and infeasible. Consider, for example, the task of learning the interaction function of the causal independence model. This task

becomes intractable when n grows; thus, reduction of the computational complexity by a factor of n can ease the problem.

In this section, we present approximations and bounds for the probability $\Pr_{\tau_k}(e^+ \mid \mathbf{c}_E)$ that can be computed in a linear number of operations with respect to the number of cause variables; and, hence, reduce the complexity by a factor of n. To illustrate the quality of the discussed techniques, we use the example noisy threshold model shown in Figure 2.2.

2.6.1 Approximations to the Poisson binomial distribution

In this section, we present two approximations to the Poisson binomial distribution that can be computed in linear time. These approximations are considered useful for very large noisy threshold models, such as discussed at the beginning of this section.

Poisson approximation

Let

$$\mathrm{P}(l; m(\mathbf{p})) = \frac{e^{-m(\mathbf{p})} m(\mathbf{p})^l}{l!}$$

denote the Poisson distribution. The following bound on the total variation distance between the Poisson binomial distribution and the Poisson distribution was established in [69]:

$$\sum_{l=0}^{\infty} |\mathrm{B}(l; \mathbf{p}) - \mathrm{P}(l; m(\mathbf{p}))| < 2 \sum_{p \in \mathbf{p}} p^2.$$

Thus, the Poisson approximation is accurate whenever the probabilistic parameters $p \in \mathbf{p}$ are small.

As an illustration, let us choose a realization $\mathbf{c}_E$ from the gastric non-Hodgkin lymphoma model such that the probabilistic parameters in $\mathbf{p}(\mathbf{c}_E)$ would be small. The Poisson binomial distribution and its Poisson approximation to compute the probability $\Pr(e^+ \mid \mathbf{c}_E)$,
$\mathbf{c}_E = \{c_1^+, c_2^-, c_3^-, c_4^+, c_5^+, c_6^+\}$ are displayed in Figure 2.5.

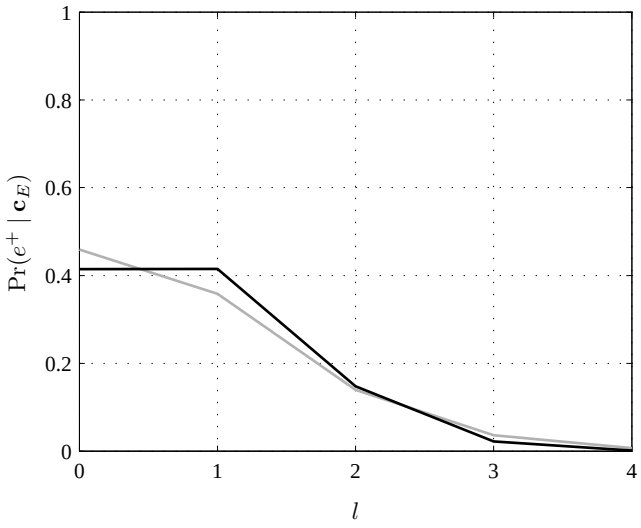


Figure 2.5: Example of the Poisson approximation (grey) to the Poisson binomial distribution (black). The number of successes is denoted by l.

Normal approximation

Another approximation to the Poisson binomial distribution reported in the literature is the approximation by the standard normal distribution [88, 111]. Let

$$\phi(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}$$

denote the normal density function and let

$$\Phi(z) = \int_{-\infty}^{z} \phi(x)dx.$$

Then, for every Poisson binomial distribution B with mean $m(\mathbf{p})$ and variance $\sigma(\mathbf{p})^2$, we have that

$$\max_{0\leq i\leq n} \left| \sum_{j=0}^{i} \mathrm{B}(j;\mathbf{p}) - \Phi\left(\frac{i-m(\mathbf{p})}{\sigma(\mathbf{p})}\right) \right| < \frac{0.7975}{\sigma(\mathbf{p})}.$$

The normal approximation is accurate when the standard deviation of the Poisson binomial distribution

$$\sigma(\mathbf{p}) = \sqrt{n\left(\mu(1-\mu) - \sigma^2\right)}$$

is large, i.e. when $n \to \infty$.

Let

$$\mathrm{N}(i; m(\mathbf{p}); \sigma(\mathbf{p})) = \Phi\left(\frac{i + \frac{1}{2} - m(\mathbf{p})}{\sigma(\mathbf{p})}\right) - \Phi\left(\frac{i - \frac{1}{2} - m(\mathbf{p})}{\sigma(\mathbf{p})}\right)$$

be a normal approximation to $\mathrm{B}(i; \mathbf{p})$. To illustrate the quality of the normal approximation, let us choose a realization $\mathbf{c}_E$ from the gastric non-Hodgkin lymphoma model with all causes being true. The Poisson binomial distribution and its normal approximation to compute the probability $\Pr(e^+ \mid \mathbf{c}_E)$, $\mathbf{c}_E = \{c_1^+, c_2^+, c_3^+, c_4^+, c_5^+, c_6^+\}$ are shown in Figure 2.6.

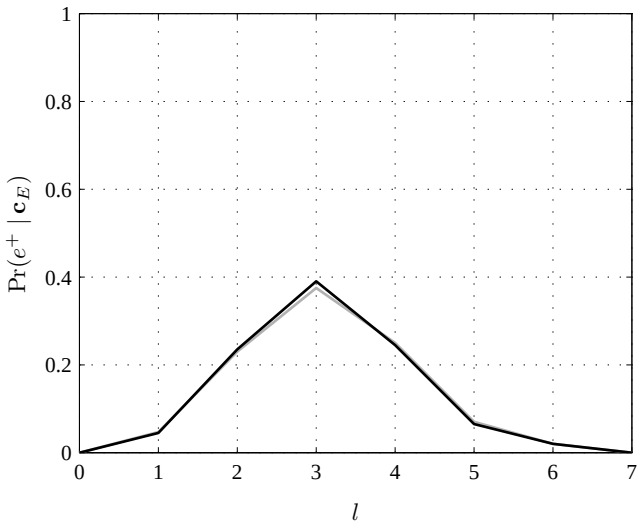


Figure 2.6: Example of the normal approximation (grey) to the Poisson binomial distribution (black). The number of successes is denoted by l.

The Poisson binomial distribution can also be approximated by the binomial distribution [90]. The binomial approximation is accurate whenever the variance σ^2 is small.

2.6.2 Bounds for the probability $\Pr_{\tau_k}(e^+ \mid \mathbf{c}_E)$

A number of bounds for the Poisson binomial distribution based on various characteristics of the probabilistic parameters $p \in \mathbf{p}$ have been reported [6, 37, 47, 53, 86]. These bounds offer another way to obtain an approximate probability $\Pr_{\tau_k}(e^+ \mid \mathbf{c}_E)$ in the noisy threshold models where n is large. All bounds for the Poisson binomial distribution concern bounds for cumulative probabilities, and hence they are well suited to bound the conditional probability of the effect in the noisy threshold models. We discuss Hoeffding's inequalities and the Percus and Percus bounds as these bounds seem to be the most suitable for the domain of the noisy threshold models.

Hoeffding's inequalities

Let $\mathrm{B}(l; m(\mathbf{p}); n)$ denote the binomial distribution. Hoeffding [47] presented the following bounds for the probabilities $\sum_{i=0}^{c} \mathrm{B}(i; \mathbf{p})$ and $\sum_{i=b}^{c} \mathrm{B}(i; \mathbf{p})$ given the mean $m(\mathbf{p})$ of the Poisson binomial distribution:

$$0 \leq \sum_{i=0}^{c} \mathrm{B}(i; \mathbf{p}) \leq \sum_{i=0}^{c} \mathrm{B}(i; m(\mathbf{p}); n) \qquad \text{if } 0 \leq c \leq m(\mathbf{p}) - 1, \tag{2.14}$$

$$\sum_{i=0}^{c} \mathrm{B}(i; m(\mathbf{p}); n) \leq \sum_{i=0}^{c} \mathrm{B}(i; \mathbf{p}) \leq 1 \qquad \text{if } m(\mathbf{p}) \leq c \leq n, \tag{2.15}$$

$$\sum_{i=b}^{c} \mathrm{B}(i; m(\mathbf{p}); n) \leq \sum_{i=b}^{c} \mathrm{B}(i; \mathbf{p}) \leq 1 \qquad \text{if } 0 \leq b \leq m(\mathbf{p}) \leq c \leq n, \tag{2.16}$$

where b and c are integers.

From Hoeffding's inequalities we obtain the following bounds for the probability $\Pr_{\tau_k}(e^+ \mid \mathbf{c}_E)$.

Proposition 5 *Let $m(\mathbf{p}(\mathbf{c}_E))$ be the mean of the probability distribution $\mathrm{B}(i; \mathbf{p}(\mathbf{c}_E))$ and let $\rho(\mathbf{c}_E) \geq k, 1 \leq k \leq n$. Then,*

- $\Pr_{\tau_k}(e^+ \mid \mathbf{c}_E) \geq \sum_{i=k}^{\rho(\mathbf{c}_E)} \mathrm{B}(i; m(\mathbf{p}(\mathbf{c}_E)); \rho(\mathbf{c}_E)) \qquad \text{if } m(\mathbf{p}(\mathbf{c}_E)) \geq k,$
- $\Pr_{\tau_k}(e^+ \mid \mathbf{c}_E) \leq \sum_{i=k}^{\rho(\mathbf{c}_E)} \mathrm{B}(i; m(\mathbf{p}(\mathbf{c}_E)); \rho(\mathbf{c}_E)) \qquad \text{if } m(\mathbf{p}(\mathbf{c}_E)) \leq k\text{–}1.$

Proof: Let $b = k$ *and* $c = \rho(\mathbf{c}_E)$*, then Inequality (2.16) becomes*

$$\sum_{i=k}^{\rho(\mathbf{c}_E)} \mathrm{B}(i; m(\mathbf{p}(\mathbf{c}_E)); \rho(\mathbf{c}_E)) \leq \sum_{i=k}^{\rho(\mathbf{c}_E)} \mathrm{B}(i; \mathbf{p}(\mathbf{c}_E)) \leq 1 \qquad \text{if } m(\mathbf{p}(\mathbf{c}_E)) \geq k.$$

Let $c = k - 1$*, then Inequality (2.15) becomes*

$$\sum_{i=0}^{k-1} \mathrm{B}(i; m(\mathbf{p}(\mathbf{c}_E)); \rho(\mathbf{c}_E)) \leq \sum_{i=0}^{k-1} \mathrm{B}(i; \mathbf{p}(\mathbf{c}_E)) \leq 1 \qquad \text{if } m(\mathbf{p}(\mathbf{c}_E)) \leq k-1.$$

Using $\sum_{i=0}^{k-1} \mathrm{B}(i; \mathbf{p}(\mathbf{c}_E)) = 1 - \sum_{i=k}^{\rho(\mathbf{c}_E)} \mathrm{B}(i; \mathbf{p}(\mathbf{c}_E))$*, we obtain*

$$0 \leq \sum_{i=k}^{\rho(\mathbf{c}_E)} \mathrm{B}(i; \mathbf{p}(\mathbf{c}_E)) \leq \sum_{i=k}^{\rho(\mathbf{c}_E)} \mathrm{B}(i; m(\mathbf{p}(\mathbf{c}_E)); \rho(\mathbf{c}_E)) \qquad \text{if } m(\mathbf{p}(\mathbf{c}_E)) \leq k-1.$$

Finally, using Proposition 2 we get

$$\Pr_{\tau_k}(e^+ \mid \mathbf{c}_E) \geq \sum_{i=k}^{\rho(\mathbf{c}_E)} \mathrm{B}(i; m(\mathbf{p}(\mathbf{c}_E)); \rho(\mathbf{c}_E)) \qquad \text{if } m(\mathbf{p}(\mathbf{c}_E)) \geq k,$$

$$\Pr_{\tau_k}(e^+ \mid \mathbf{c}_E) \leq \sum_{i=k}^{\rho(\mathbf{c}_E)} \mathrm{B}(i; m(\mathbf{p}(\mathbf{c}_E)); \rho(\mathbf{c}_E)) \qquad \text{if } m(\mathbf{p}(\mathbf{c}_E)) \leq k-1.$$

■

Since $m(\mathbf{p}(\mathbf{c}_E))$ can be computed in linear time, the Hoeffding's bounds can be computed in linear time as well.

We examined the tightness of the Hoeffding's bounds for the probability $\Pr_{\tau_k}(e^+ \mid \mathbf{c}_E)$ in the non-Hodgkin lymphoma model. The results for all possible realizations of $\mathbf{c}_E, E = \{1, \ldots, 6\}$ are summarized as the difference between the bounds and presented in Figure 2.7. The probability for 32 realizations of $\mathbf{c}_E$ could not be bounded as their $m(\mathbf{p}(\mathbf{c}_E))$ value falls into the interval $(1, 2)$. The conditional probability for the other realizations was bounded by intervals smaller than 0.4.

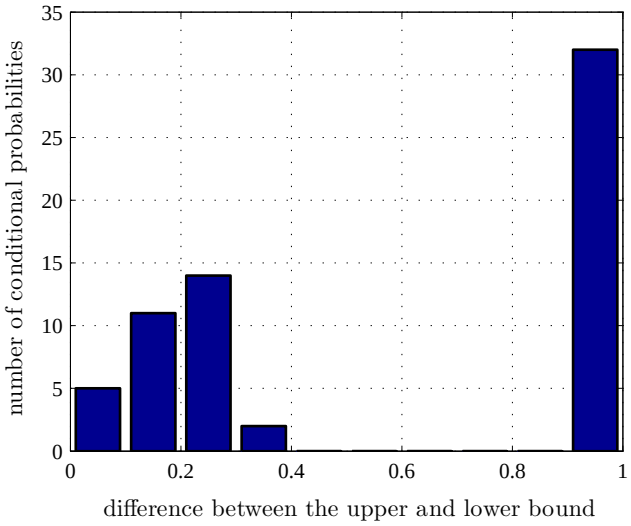


Figure 2.7: Hoeffding's bounds for the conditional probability $\Pr_{\tau_k}(e^+ \mid \mathbf{c}_E)$ in the non-Hodgkin lymphoma model.

Percus and Percus bounds

Percus and Percus introduced a lower bound for the sum of the Poisson binomial probabilities $\sum_{i=0}^{c} \mathrm{B}(i;\mathbf{p})$. To compute the bound, the probability $\mathrm{B}(0;\mathbf{p})$ is used [86]:

$$\sum_{i=0}^{c} \mathrm{B}(i;\mathbf{p}) \geq \mathrm{B}(0;\mathbf{p}) \sum_{i=0}^{c} \binom{n}{i} \left((\mathrm{B}(0;\mathbf{p}))^{-\frac{1}{n}} - 1 \right)^{i}. \tag{2.17}$$

From Equation (2.17), we derive an upper bound for the sum of the Poisson binomial probabilities $\sum_{i=0}^{c} \mathrm{B}(i;\mathbf{p})$. Let $\mathbf{q} = (1-p_1, \ldots, 1-p_n)$ denote the probabilities of failures in the Poisson trials. Then, the relation between $\mathrm{B}(i;\mathbf{p})$ and $\mathrm{B}(i;\mathbf{q})$ is

$$\sum_{i=0}^{c} \mathrm{B}(i;\mathbf{p}) = \sum_{i=n-c}^{n} \mathrm{B}(i;\mathbf{q}) = 1 - \sum_{i=0}^{n-c-1} \mathrm{B}(i;\mathbf{q}). \tag{2.18}$$

We can rewrite Equation (2.17) as

$$\begin{aligned}\sum_{i=0}^{c} \mathrm{B}(i;\mathbf{q}) &\geq \mathrm{B}(0;\mathbf{q}) \sum_{i=0}^{c} \binom{n}{i} \left((\mathrm{B}(0;\mathbf{q}))^{-\frac{1}{n}} - 1 \right)^i \\ &= \mathrm{B}(n;\mathbf{p}) \sum_{i=0}^{c} \binom{n}{i} \left((\mathrm{B}(n;\mathbf{p}))^{-\frac{1}{n}} - 1 \right)^i . \end{aligned} \tag{2.19}$$

Finally, combining (2.18) and (2.19), we obtain an upper bound for the sum of the Poisson binomial probabilities, given by

$$\sum_{i=0}^{c} \mathrm{B}(i;\mathbf{p}) \leq 1 - \mathrm{B}(n;\mathbf{p}) \sum_{i=0}^{n-c-1} \binom{n}{i} \left((\mathrm{B}(n;\mathbf{p}))^{-\frac{1}{n}} - 1 \right)^i . \tag{2.20}$$

Using the Percus and Percus bounds, we find the following upper and lower bounds for the probability $\Pr_{\tau_k}(e^+ \mid \mathbf{c}_E)$.

Proposition 6 *Let* $\rho(\mathbf{c}_E) \geq k, 1 \leq k \leq n$, *then*

- $\Pr_{\tau_k}(e^+ \mid \mathbf{c}_E) \leq 1 - \mathrm{B}(0;\mathbf{p}(\mathbf{c}_E)) \sum_{i=0}^{k-1} \binom{\rho(\mathbf{c}_E)}{i} \left((\mathrm{B}(0;\mathbf{p}(\mathbf{c}_E)))^{-\frac{1}{\rho(\mathbf{c}_E)}} - 1 \right)^i$,
- $\Pr_{\tau_k}(e^+ \mid \mathbf{c}_E) \geq \mathrm{B}(\rho(\mathbf{c}_E);\mathbf{p}(\mathbf{c}_E)) \sum_{i=0}^{\rho(\mathbf{c}_E)-k} \binom{\rho(\mathbf{c}_E)}{i} \left((\mathrm{B}(\rho(\mathbf{c}_E);\mathbf{p}(\mathbf{c}_E)))^{-\frac{1}{\rho(\mathbf{c}_E)}} - 1 \right)^i$.

Proof: Let $c = k - 1$. *Then, using Proposition 2, inequalities (2.17) and (2.20) become*

$$\begin{aligned}\Pr_{\tau_k}(e^+ \mid \mathbf{c}_E) &= 1 - \sum_{i=0}^{k-1} \mathrm{B}(i;\mathbf{p}(\mathbf{c}_E)) \\ &\leq 1 - \mathrm{B}(0;\mathbf{p}(\mathbf{c}_E)) \sum_{i=0}^{k-1} \binom{\rho(\mathbf{c}_E)}{i} \left((\mathrm{B}(0;\mathbf{p}(\mathbf{c}_E)))^{-\frac{1}{\rho(\mathbf{c}_E)}} - 1 \right)^i , \\ \Pr_{\tau_k}(e^+ \mid \mathbf{c}_E) &= 1 - \sum_{i=0}^{k-1} \mathrm{B}(i;\mathbf{p}(\mathbf{c}_E)) \\ &\geq \mathrm{B}(\rho(\mathbf{c}_E);\mathbf{p}(\mathbf{c}_E)) \sum_{i=0}^{\rho(\mathbf{c}_E)-k} \binom{\rho(\mathbf{c}_E)}{i} \left((\mathrm{B}(\rho(\mathbf{c}_E);\mathbf{p}(\mathbf{c}_E)))^{-\frac{1}{\rho(\mathbf{c}_E)}} - 1 \right)^i . \end{aligned}$$

■

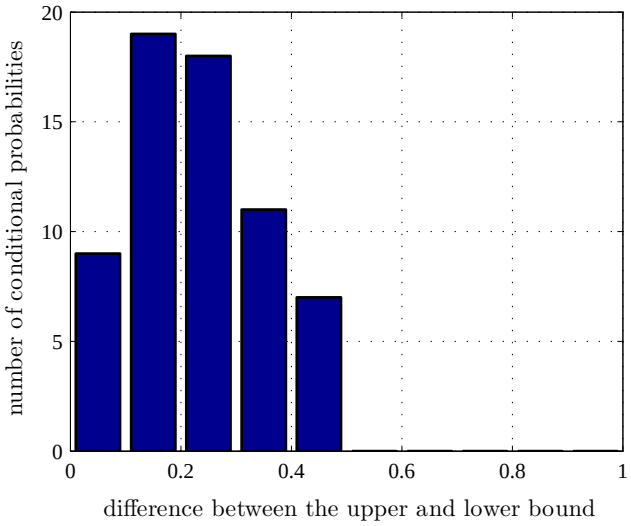


Figure 2.8: Percus and Percus bounds for the probability $\Pr_{\tau_k}(e^+ \mid \mathbf{c}_E)$ in the non-Hodgkin lymphoma model.

Since $\mathrm{B}(0; \mathbf{p}(\mathbf{c}_E))$ and $\mathrm{B}(\rho(\mathbf{c}_E); \mathbf{p}(\mathbf{c}_E))$ can be computed in linear time, the Percus and Percus bounds can be computed in linear time as well.

We have examined the tightness of the Percus and Percus bounds for the probability $\Pr_{\tau_k}(e^+ \mid \mathbf{c}_E)$ in the non-Hodgkin lymphoma model. The results for all possible realizations of $\mathbf{c}_E, E = \{1, \ldots, 6\}$ are summarized as the difference between the bounds and presented in Figure 2.8.

The best bounds for the probability $\Pr_{\tau_k}(e^+ \mid \mathbf{c}_E)$ can be achieved by combining Hoeffding's and Percus and Percus bounds. In our simulations, non-trivial Hoeffding's bounds were always at least as tight as the Percus and Percus bounds. See Figure 2.9 for the results.

2.7 Discussion

In this chapter, we discussed the inference problem in causal independence models based on the Boolean threshold functions. We showed that the inference problem in causal independence models based on symmetric Boolean functions boils down to computing the conditional probability of the effect. We established a connection between the conditional probabil-

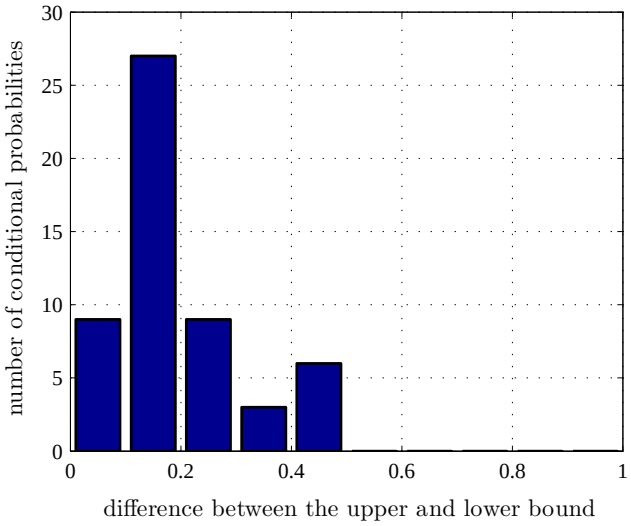


Figure 2.9: Combined Hoeffding's and Percus and Percus bounds for the probability $\Pr_{\tau_k}(e^+ \mid \mathbf{c}_E)$ in the non-Hodgkin lymphoma model.

ity of the effect in the noisy threshold models and the Poisson binomial distribution. We investigated how the properties of the Poisson binomial distribution can be used to compute both exact and approximate probability values for the effect variable. The approximate methods are expected to be useful in very large noisy threshold models.

Even though this chapter has focused on the inference in the noisy threshold models, most of the presented theory can be used as a basis for the inference in causal independence models with any symmetric Boolean function defining interaction among causes.

Chapter 3

Parameter Learning

In this chapter, we study the problem of learning the parameters in symmetric causal independence models. We present a computationally efficient expectation-maximization (EM) algorithm to learn parameters in these models, where the computational scheme of the Poisson binomial distribution is used to compute the conditional probabilities in the E-step. We study the computational complexity and convergence of the developed algorithm. The presented EM algorithm allows us to assess the practical usefulness of symmetric causal independence models. In the assessment, the models are applied to a classification task; they perform competitively with state-of-the-art classifiers.

This chapter is based on: R. Jurgelenaite and T. Heskes. Learning symmetric causal independence models. *Machine Learning*, 2008.

3.1 Introduction

Even though in some real-world problems the intermediate variables in causal independence models are observable [116], in many problems these variables are latent. In such problems, the conditional probability distribution of the effect given the causes depends on unknown parameters, which have to be estimated from data using *maximum likelihood* (ML) or

maximum a posteriori (MAP) methods. One of the most widespread techniques for finding ML or MAP estimates is the *expectation-maximization* (EM) method. A direct application of the EM method to learn parameters in symmetric causal independence models is not tractable for models with many causes as the method is exponential in the number of causes. At least two variants of the EM method that make use of specific properties of causal independence models can be found in literature. Meek and Heckerman [76] provided a general algorithm to use the EM method to compute the maximum likelihood estimate of the parameters in causal interaction models, which are a generalization of noisy-max and noisy-additive models that allows a more flexible model structure. Vomlel [118] described the application of an EM algorithm to learn the parameters in the noisy OR model. However, the proposed schemes of the EM algorithms are either too abstract or too specific to be directly applied to the general case of parameter learning in causal independence models.

Learning the parameters in causal independence models with a symmetric Boolean function as an interaction function, further referred to as symmetric causal independence models, is the main topic of this chapter. We develop an EM algorithm to learn the parameters in symmetric causal independence models, and study computational complexity and convergence of the algorithm. The EM algorithm enables us to assess the practical usefulness of these extended models. In the assessment, we use symmetric causal independence models as classifiers. Experimental results show the competitive performance of symmetric causal independence models compared to the noisy OR model as well as other widely-used classifiers.

The remainder of this chapter is organised as follows. In the following section, we establish the relationship between the Poisson binomial probabilities and the conditional probability distribution given the Boolean exact function. In Section 3.3, we first describe the general scheme of the EM algorithm and then develop the EM algorithm for parameter learning in symmetric causal independence models. The maxima of the log-likelihood function for the symmetric causal independence models are examined in Section 3.4. Finally, Section 3.5 presents the experimental results, and conclusions are drawn in Section 3.6. The Appendix contains a number of properties that enable a reduction of computational complexity of the E-step of the EM algorithm.

3.2 The Poisson binomial distribution and symmetric causal independence models

Using the property that symmetric Boolean functions can be decomposed in terms of exact Boolean functions (see Equation 2.2), the conditional probability of the occurrence of the effect E given the causes $C_1, \ldots, C_n$ can be decomposed in terms of probabilities that exactly l hidden variables are true as

$$\Pr(e^+ \mid c_1, \ldots, c_n) = \sum_{\substack{0 \leq l \leq n \\ \gamma_l}} \sum_{\epsilon_l(h_1,\ldots,h_n)=\top} \prod_{i=1}^{n} \Pr(h_i \mid c_i). \tag{3.1}$$

As we will show next, the conditional probability of the occurrence of the effect E given the causes $C_1, \ldots, C_n$ in symmetric causal independence models is closely related to the Poisson binomial distribution.

Let l denote the number of successes in n independent trials, where p_i is the probability of success in the ith trial, $i = 1, \ldots, n$; let $\mathbf{p} = (p_1, \ldots, p_n)$, then $\mathrm{B}(l; \mathbf{p})$ denotes the *Poisson binomial distribution* [30, 69]:

$$\mathrm{B}(l; \mathbf{p}) = \left\{ \prod_{i=1}^{n} (1 - p_i) \right\} \sum_{1 \leq j_1 < \ldots < j_l \leq n} \prod_{z=1}^{l} \frac{p_{j_z}}{1 - p_{j_z}}. \tag{3.2}$$

To put it in words, the Poisson binomial distribution is a discrete probability distribution of the number of successes in a sequence of n independent experiments, each of which yields success with a different probability. When all trials have the same probability of success p, $\mathrm{B}(l; \mathbf{p})$ reduces to the binomial distribution: $\mathrm{B}(l; p) = \binom{n}{l} p^l (1-p)^{n-l}$.

Let us define a vector of probabilistic parameters $\mathbf{p}(c_1, \ldots, c_n) = (p_1, \ldots, p_n)$ with $p_i = \Pr(h_i^+ \mid c_i)$. The relationship between the Poisson binomial probabilities and the conditional probability distribution given the Boolean exact function is as stated in the following proposition.

Proposition 7 *It holds that:*

$$\sum_{\epsilon_l(h_1,\ldots,h_n)=\top} \prod_{i=1}^{n} \Pr(h_i \mid c_i) = \mathrm{B}(l; \mathbf{p}(c_1, \ldots, c_n)). \tag{3.3}$$

Proof: Note that in Equation (3.1), the sum

$$\sum_{\epsilon_l(h_1,\ldots,h_n)=\top} \prod_{i=1}^{n} \Pr(h_i \mid c_i)$$

was defined as the probability that exactly l hidden variables are true. A hidden variable H_i can be seen as an independent trial whose probability of success is $\Pr(h_i^+ \mid c_i)$. Combining the definition of the vector of probabilistic parameters $\mathbf{p}(c_1, \ldots, c_n)$ and the definition of the Poisson binomial distribution, the result in the premise of this proposition is obtained. ■

Now, we can establish the relationship between the conditional probability of the effect given the causes in a symmetric causal independence model and the Poisson binomial distribution.

Proposition 8 *It holds that:*

$$\Pr(e^+ \mid c_1, \ldots, c_n) = \sum_{i=0}^{n} \mathrm{B}(i; \mathbf{p}(c_1, \ldots, c_n))\gamma_i.$$

Proof: The proof follows from Equation (3.1) and Proposition 7. ■

The relationship described in Proposition 8 allows us to use the theory of the well-studied Poisson binomial distribution in the context of symmetric causal independence models. The properties of the Poisson binomial distribution will be of major importance in developing a computationally efficient EM algorithm for symmetric causal independence models.

3.3 The EM algorithm

For a symmetric causal independence model to be complete, its structure, interaction function and parameters need to be determined. The structure of symmetric causal independence models is fixed, and the interaction function f is assumed to be known. To have a fully specified symmetric causal independence model, we need to estimate the unknown parameters in the model, i.e. the parameters of the conditional distributions of hidden variables $H_1, \ldots, H_n$. Given the assumption that absent causes do not contribute to the effect, the unknown parameters in the model are $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ where $\theta_i = \Pr(h_i^+ \mid c_i^+)$.

3.3.1 Maximum likelihood estimate and basic EM

Let $\mathbf{D} = \{\mathbf{x}^1, \ldots, \mathbf{x}^N\}$ be a data set of independent and identically distributed settings of the observed variables in a symmetric causal independence model where

$$\mathbf{x}^j = (\mathbf{c}^j, e^j) = (c_1^j, \ldots, c_n^j, e^j).$$

We focus on the problem of estimating the parameters when no additional information about the model is available, i.e. we do not have any prior knowledge about the parameters $\boldsymbol{\theta}$. To learn $\boldsymbol{\theta}$, we maximize the conditional log-likelihood

$$CLL(\boldsymbol{\theta}) = \sum_{j=1}^{N} \ln \Pr(e^j \mid \mathbf{c}^j, \boldsymbol{\theta}).$$

The value of $\boldsymbol{\theta}$ which maximizes the conditional log-likelihood is known as the *maximum likelihood estimate* for $\boldsymbol{\theta}$. Expectation-maximization (EM) [26] is a general method to find the maximum likelihood estimate of the parameters in probabilistic models, where the data is incomplete or the model has hidden variables.

The EM method can be explained and derived in several ways. We derive an EM algorithm on the basis of the explanation of EM in terms of lower-bound maximization [81]. We start from the following simple identity:

$$\ln \Pr(e^j \mid \mathbf{c}^j, \boldsymbol{\theta}) = \ln \Pr(\mathbf{h}, e^j \mid \mathbf{c}^j, \boldsymbol{\theta}) - \ln \Pr(\mathbf{h} \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}) \tag{3.4}$$

and take expectations of both sides, treating $\mathbf{H}$ as a random variable with the distribution $\Pr(\mathbf{h} \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(old)})$ where $\boldsymbol{\theta}^{(old)}$ is the current (old) estimate. The left hand side of Equation (3.4) does not depend on $\mathbf{H}$, so averaging over $\mathbf{H}$ yields

$$\begin{aligned} \ln \Pr(e^j \mid \mathbf{c}^j, \boldsymbol{\theta}) &= \sum_{\mathbf{h}} \Pr(\mathbf{h} \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(old)}) \ln \Pr(\mathbf{h}, e^j \mid \mathbf{c}^j, \boldsymbol{\theta}) \\ &- \sum_{\mathbf{h}} \Pr(\mathbf{h} \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(old)}) \ln \Pr(\mathbf{h} \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}). \end{aligned} \tag{3.5}$$

The key result for the EM algorithm is that the last term in the above equation is maximized at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(old)}$, thus any increase of the first term on the right side of Equation (3.5) is guaranteed to increase the expected complete (conditional) log-likelihood.

Let us denote

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(z)}) = \sum_{j=1}^{N} \sum_{\mathbf{h}} \Pr(\mathbf{h} \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)}) \ln \Pr(\mathbf{h}, e^j \mid \mathbf{c}^j, \boldsymbol{\theta}). \tag{3.6}$$

The EM algorithm at each iteration maximizes the functional

$$\boldsymbol{\theta}^{(z+1)} = \operatorname*{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(z)}).$$

3.3.2 Maximization step

To maximize $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(z)})$, we need to compute the partial derivatives of this functional with respect to each parameter, set them equal to zero and solve the system of equations. We start by transforming $\ln \Pr(\mathbf{h}, e^j \mid \mathbf{c}^j, \boldsymbol{\theta})$ so that it becomes a sum of logarithms:

$$\begin{aligned} \ln \Pr(\mathbf{h}, e^j \mid \mathbf{c}^j, \boldsymbol{\theta}) &= \ln \left(\Pr(e^j \mid \mathbf{h}) \prod_{i=1}^{n} \Pr(h_i \mid c_i^j, \theta_i) \right) \\ &= \ln \Pr(e^j \mid \mathbf{h}) + \sum_{i=1}^{n} \ln \Pr(h_i \mid c_i^j, \theta_i). \end{aligned} \tag{3.7}$$

The conditional probability $\Pr(h_i \mid c_i^j, \theta_i)$ can be written in the form

$$\Pr(h_i \mid c_i^j, \theta_i) = c_i^j h_i \theta_i + c_i^j (1 - h_i)(1 - \theta_i) + (1 - c_i^j)(1 - h_i). \tag{3.8}$$

Combining (3.6), (3.7) and (3.8), we obtain

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(z)}) = \sum_{j=1}^{N} \sum_{\mathbf{h}} \Pr(\mathbf{h} \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)}) \left(\ln \Pr(e^j \mid \mathbf{h}) + \sum_{i=1}^{n} \ln \left(\theta_i c_i^j (2h_i - 1) + 1 - h_i \right) \right).$$

Then taking the partial derivatives with respect to each parameter θ_k and setting them equal to zero yields

$$\frac{\partial Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(z)})}{\partial \theta_k} = \sum_{j=1}^{N} \sum_{\mathbf{h}} \Pr(\mathbf{h} \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)}) \frac{c_k^j (2h_k - 1)}{\theta_k c_k^j (2h_k - 1) + 1 - h_k} = 0. \tag{3.9}$$

Now let us define $\mathbf{h}_{\setminus k} = \{h_1, \ldots, h_{k-1}, h_{k+1}, \ldots, h_n\}$. Equation (3.9) can be simplified by writing it as a sum over the states of the hidden variable H_k:

$$\sum_{j=1}^{N} c_k^j \sum_{\mathbf{h}_{\setminus k}} \left(\Pr(\mathbf{h}_{\setminus k}, h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)}) \frac{1}{\theta_k c_k^j} - \Pr(\mathbf{h}_{\setminus k}, h_k^- \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)}) \frac{1}{1 - \theta_k c_k^j} \right) = 0. \tag{3.10}$$

It can be shown that Equation (3.10) is solved by

$$\begin{aligned} \theta_k &= \frac{\sum_{j=1}^{N} \sum_{\mathbf{h}_{\setminus k}} \Pr(\mathbf{h}_{\setminus k}, h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)})}{\sum_{j=1}^{N} c_k^j \sum_{\mathbf{h}_{\setminus k}} \left(\Pr(\mathbf{h}_{\setminus k}, h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)}) + \Pr(\mathbf{h}_{\setminus k}, h_k^- \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)}) \right)} \\ &= \frac{\sum_{j=1}^{N} \Pr(h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)})}{\sum_{j=1}^{N} c_k^j}. \end{aligned} \tag{3.11}$$

In the next section, we derive the expectation step, which corresponds to computing the conditional probabilities

$$\Pr(h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)})$$

for all $k = 1, \ldots, n,\ j = 1, \ldots, N$ where $c_k^j = 1$.

3.3.3 Expectation step

Using Bayes' rule, we can write the conditional probability of $\mathbf{H}$ given the data sample $\mathbf{x}^j$ and the parameters $\boldsymbol{\theta}^{(z)}$ as

$$\Pr(\mathbf{h} \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)}) = \frac{\Pr(e^j \mid \mathbf{h}) \prod_{i=1}^{n} \Pr(h_i \mid c_i^j, \boldsymbol{\theta}^{(z)})}{\Pr(e^j \mid \mathbf{c}^j, \boldsymbol{\theta}^{(z)})}.$$

By marginalizing $\mathbf{h}_{\setminus k}$ out we obtain the conditional probability of the hidden variable H_k being true:

$$\Pr(h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)}) = \frac{\Pr(h_k^+ \mid c_k^j, \theta_k^{(z)})}{\Pr(e^j \mid \mathbf{c}^j, \boldsymbol{\theta}^{(z)})} \sum_{\mathbf{h}_{\setminus k}} \Pr(e^j \mid \mathbf{h}_{\setminus k}, h_k^+) \prod_{\substack{1 \leq i \leq n \\ i \neq k}} \Pr(h_i \mid c_i^j, \theta_i^{(z)}). \tag{3.12}$$

Computing the probability $\Pr(h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)})$ from Equation (3.12) in a straightforward way requires summing over 2^n terms, which is computationally expensive. In [15, 49] it was shown that using recursive methods the Poisson binomial distribution can be computed in a quadratic number of operations with respect to n. Therefore, expressing Equation (3.12) in terms of the Poisson binomial probabilities is an obvious way to reduce the computational cost of the algorithm.

Let us define $\hat{\boldsymbol{\theta}}_{(k=1)}^{(z)} = (\hat{\theta}_1^{(z)}, \ldots, \hat{\theta}_n^{(z)})$ where

$$\hat{\theta}_k^{(z)} = 1 \text{ and } \hat{\theta}_i^{(z)} = \theta_i^{(z)}, \ \forall i \neq k.$$

Using the defined vector $\hat{\boldsymbol{\theta}}_{(k=1)}^{(z)}$ and $\Pr(h_k^+ \mid c_k^j, \theta_k^{(z)}) = c_k^j \theta_k^{(z)}$, Equation (3.12) takes the form

$$\Pr(h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)}) = \frac{c_k^j \theta_k^{(z)} \Pr(e^j \mid \mathbf{c}^j, \hat{\boldsymbol{\theta}}_{(k=1)}^{(z)})}{\Pr(e^j \mid \mathbf{c}^j, \boldsymbol{\theta}^{(z)})}. \tag{3.13}$$

Now, we can express the obtained result in terms of the Poisson binomial probabilities. First, let us define

$$\begin{aligned} &\mathbf{p}^{(z,j)} = (p_1^{(z,j)}, \ldots, p_n^{(z,j)}) \quad \text{where} \quad p_i^{(z,j)} = \theta_i^{(z)} c_i^j, \\ &\hat{\mathbf{p}}_{(k=1)}^{(z,j)} = (\hat{p}_1^{(z,j)}, \ldots, \hat{p}_n^{(z,j)}) \quad \text{where} \quad \hat{p}_k = 1 \text{ and } \hat{p}_i^{(z,j)} = \theta_i^{(z)} c_i^j, \forall_{i \neq k}. \end{aligned}$$

From the property of the Poisson binomial distribution [22]

$$\mathrm{B}(i; \mathbf{p}) = \mathrm{B}(i; \mathbf{p}_{\setminus k})(1 - p_k) + \mathrm{B}(i-1; \mathbf{p}_{\setminus k}) p_k \tag{3.14}$$

it follows that

$$\mathrm{B}(i; \hat{\mathbf{p}}_{(k=1)}^{(z,j)}) = \mathrm{B}(i-1; \mathbf{p}_{\setminus k}^{(z,j)}). \tag{3.15}$$

Using the identity (3.15) and Proposition 8, the left hand side of (3.13) can be expressed in terms of the Poisson binomial probabilities as

$$\Pr(h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)}) = \begin{cases} \dfrac{p_k^{(z,j)} \sum_{i=0}^{n-1} \mathrm{B}\left(i; \mathbf{p}_{\setminus k}^{(z,j)}\right) \gamma_{i+1}}{\sum_{i=0}^{n} \mathrm{B}\left(i; \mathbf{p}^{(z,j)}\right) \gamma_i} & \text{if } e^j = 1 \\ \dfrac{p_k^{(z,j)} \left(1 - \sum_{i=0}^{n-1} \mathrm{B}\left(i; \mathbf{p}_{\setminus k}^{(z,j)}\right) \gamma_{i+1}\right)}{1 - \sum_{i=0}^{n} \mathrm{B}\left(i; \mathbf{p}^{(z,j)}\right) \gamma_i} & \text{if } e^j = 0 \end{cases} \tag{3.16}$$

Summarizing, the $(z+1)$-th iteration of the EM algorithm for symmetric causal independence models is given by:

Expectation step: For every instance $\mathbf{x}^j = (\mathbf{c}^j, e^j)$ with $j = 1, \ldots, N$, we form

$$\mathbf{p}^{(z,j)} = (p_1^{(z,j)}, \ldots, p_n^{(z,j)}) \quad \text{where} \quad p_i^{(z,j)} = \theta_i^{(z)} c_i^j.$$

Subsequently, the probability $\mathrm{P}(h_k^+ \mid \mathbf{c}^j, e^j, \boldsymbol{\theta}^{(z)})$ is computed from (3.16) for each hidden variable $H_k, k = 1, \ldots, n$.

Maximization step: Update the parameter estimates using Equation (3.11).

The expectation and maximization steps are repeated until some convergence criterion is satisfied.

3.3.4 Computational complexity of the expectation step

The computation of the conditional probabilities in the E-step is a probabilistic inference task; therefore, it can be performed by an inference algorithm. One way to perform efficient inference in causal independence models is first to transform the models using parent divorcing [83] or temporal transformation [42] techniques and then to apply an exact inference algorithm. Zagorecki et al. [126] used the latter technique to transform probabilistic causal independence models (models where the combination function does not need to be deterministic) to make inference efficient and

applied the standard EM algorithm to learn the parameters. Another way to perform efficient inference in causal independence models is the VE_1 algorithm [128], which factorizes the conditional probabilities into a combination of smaller factors to obtain a finer-grain factorization of the joint probability distribution and makes use of this factorization. Using one of the transformation techniques or the VE_1 algorithm, the conditional probability $\Pr(h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)})$ can be computed in a quadratic number of operations with respect to n; however, there was no investigation how these methods can be used for parameter learning in causal independence models.

The Poisson binomial distribution and, consequently, the conditional probability $\Pr(h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)})$ can be computed in quadratic time with respect to n using recursive methods, and estimated in linear time with respect to n using approximation and bounding techniques (see Chapter 2 for a review of these techniques). A naive computation of the exact probabilities $\Pr(h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)})$ in the expectation step requires $O(n^3)$ operations for every data instance $\mathbf{x}^j = (e^j, \mathbf{c}^j)$ at every iteration of the algorithm. For a problem with a large number of causes n, this cubic complexity can become a computational bottleneck. Using the theory of the Poisson binomial distribution, however, the computational complexity can be reduced to $O\left(n_*^2\right)$ operations, where n_* is the number of probabilistic parameters $p_i^{(z,j)}, i = 1, \ldots, n$ such that $0 < p_i^{(z,j)} < 1$. See the Appendix for details.

3.4 Analysis of the maxima of the log-likelihood function

Generally, there is no guarantee that the EM algorithm will converge to a global maximum of the conditional log-likelihood. In this section, we investigate the maxima of the conditional log-likelihood function for symmetric causal independence models.

3.4.1 Noisy OR and noisy AND models

In this section, we show that the conditional log-likelihood for the noisy OR and the noisy AND models has only global maxima. Since the conditional log-likelihood function for these models is not necessarily concave, we use a monotonic transformation to prove the absence of stationary points other

than global maxima.

First, we establish a connection between the maxima of the log-likelihood function and the maxima of the corresponding composite function.

Proposition 9 ***(Global optimality condition for concave functions [8])***
Suppose $h(\mathbf{q}) : \mathbf{Q} \rightarrow \Re$ *is concave and differentiable on* $\mathbf{Q}$*. Then,* $\mathbf{q}^* \in \mathbf{Q}$ *is a global maximum if and only if*

$$\nabla h(\mathbf{q}^*) = \left(\frac{\partial h(\mathbf{q}^*)}{\partial q_1}, \ldots, \frac{\partial h(\mathbf{q}^*)}{\partial q_n} \right)^T = \mathbf{0}.$$

Further, we consider the function

$$CLL(\boldsymbol{\theta}) = h(\mathbf{q}(\boldsymbol{\theta})).$$

Let $CLL(\boldsymbol{\theta})$ and $h(\mathbf{q}(\boldsymbol{\theta}))$ be twice differentiable functions, and let $\mathbf{q}(\boldsymbol{\theta})$ be a differentiable, injective function where $\boldsymbol{\theta}(\mathbf{q})$ is its inverse. Then, the following relationship between the stationary points of the functions CLL and h holds.

Lemma 10 *Suppose* $\boldsymbol{\theta}^*$ *is a stationary point of* $CLL(\boldsymbol{\theta})$*. Then, there is a corresponding point* $\mathbf{q}(\boldsymbol{\theta}^*)$ *which is a stationary point of* $h(\mathbf{q}(\boldsymbol{\theta}))$*.*

Proof: Since the function $\mathbf{q}(\boldsymbol{\theta})$ *is differentiable and injective, its Jacobian matrix* $\frac{\partial(q_1,\ldots,q_n)}{\partial(\theta_1,\ldots,\theta_n)}$ *is positive definite. Therefore, from the chain rule it follows that if* $\nabla CLL(\boldsymbol{\theta}^*) = \mathbf{0}$*, then* $\nabla h(\mathbf{q}(\boldsymbol{\theta}^*)) = \mathbf{0}$*.* ■

Proposition 11 *If* $h(\mathbf{q}(\boldsymbol{\theta}))$ *is concave and* $\boldsymbol{\theta}^*$ *is a stationary point of* $CLL(\boldsymbol{\theta})$*, then* $\boldsymbol{\theta}^*$ *is a global maximum.*

Proof: If $\boldsymbol{\theta}^*$ *is a stationary point, then from Lemma 10 it follows that* $\mathbf{q}(\boldsymbol{\theta}^*)$ *is also stationary. From the global optimality condition for concave functions, the stationary point* $\mathbf{q}(\boldsymbol{\theta}^*)$ *is a maximum of* $h(\mathbf{q}(\boldsymbol{\theta}))$*; thus, from the definition of a global maximum, we get that for all* $\boldsymbol{\theta}$

$$CLL(\boldsymbol{\theta}) = h(\mathbf{q}(\boldsymbol{\theta})) \leq h(\mathbf{q}(\boldsymbol{\theta}^*)) = CLL(\boldsymbol{\theta}^*).$$

■

Given Proposition 11, the absence of local optima can be proven by introducing a monotonic transformation $\mathbf{q}(\boldsymbol{\theta})$ such that the composite function $h(\mathbf{q}(\boldsymbol{\theta}))$ would be concave. It is well known that the log-likelihood function for logistic regression is concave, i.e. has no local optima. We will use transformations that allow us to write the log-likelihood for the noisy OR and the noisy AND models in a similar form as that of the logistic regression model.

The conditional probability of the effect in the noisy OR model can be written as

$$\begin{aligned}\Pr(e^+ \mid \mathbf{c}, \boldsymbol{\theta}) &= 1 - \prod_{i=1}^{n} \Pr(h_i^- \mid c_i) = 1 - \prod_{i=1}^{n} (1-\theta_i)^{c_i} \\ &= 1 - \exp\left(\sum_{i=1}^{n} \ln(1-\theta_i) c_i\right).\end{aligned}$$

Let us choose a monotonic transformation $q_i = -\ln(1-\theta_i), i = 1, \ldots, n$. Then, the conditional probability of the effect in the noisy OR model equals

$$\Pr(e^+ \mid \mathbf{c}, \mathbf{q}) = 1 - e^{-\mathbf{q}^T \mathbf{c}}.$$

Let us define $z^j = \mathbf{q}^T \mathbf{c}^j$ and $f(z^j) = \Pr(e^+ \mid \mathbf{c}^j, \mathbf{q})$, then the function h reads

$$h(\mathbf{q}) = \sum_{j=1}^{N} e^j \ln f(z^j) + (1 - e^j) \ln(1 - f(z^j)). \tag{3.17}$$

Since $f'(z^j) = 1 - f(z^j)$, the first derivative of h is

$$\frac{\partial h(\mathbf{q})}{\partial \mathbf{q}} = \sum_{j=1}^{N} \frac{f'(z^j)(e^j - f(z^j))}{f(z^j)(1 - f(z^j))} \mathbf{c}^j = \sum_{j=1}^{N} \frac{e^j - f(z^j)}{f(z^j)} \mathbf{c}^j.$$

To prove that the function h is concave, we need to prove that its Hessian matrix is negative semidefinite. The Hessian matrix of h reads

$$\frac{\partial^2 h(\mathbf{q})}{\partial \mathbf{q} \partial \mathbf{q}^T} = -\sum_{j=1}^{N} \frac{1 - f(z^j)}{f(z^j)^2} e^j \mathbf{c}^j \mathbf{c}^{j\,T} \leq 0.$$

As the Hessian matrix of h is negative semidefinite, the function h is concave. Therefore, from Proposition 11 it follows that every stationary point of the log-likelihood function for the noisy OR model is a global maximum.

The conditional probability of the effect in the noisy AND model can be written as:

$$\Pr(e^+ \mid \mathbf{c}, \boldsymbol{\theta}) = \prod_{i=1}^{n} \Pr(h_i^+ \mid c_i) = \prod_{i=1}^{n} \theta_i^{c_i} = \exp\left(\sum_{i=1}^{n} \ln \theta_i c_i\right)$$

Let us choose a monotonic transformation $q_i = \ln \theta_i, i = 1, \ldots, n$. Then the conditional probability of the effect in the noisy AND model equals

$$\Pr(e^+ \mid \mathbf{c}, \mathbf{q}) = e^{\mathbf{q}^T \mathbf{c}}$$

Let us define $z^j = \mathbf{q}^T \mathbf{c}^j$ and $f(z^j) = \Pr(e^+ \mid \mathbf{c}^j, \mathbf{q})$. The function h is the same as for the noisy OR model in Equation (3.17). Combined with $f'(z^j) = f(z^j)$, it yields the first derivative of h

$$\frac{\partial h(\mathbf{q})}{\partial \mathbf{q}} = \sum_{j=1}^{N} \frac{f'(z^j)(e^j - f(z^j))}{f(z^j)(1 - f(z^j))} \mathbf{c}^j = \sum_{j=1}^{N} \frac{e^j - f(z^j)}{1 - f(z^j)} \mathbf{c}^j$$

and Hessian matrix

$$\frac{\partial^2 h(\mathbf{q})}{\partial \mathbf{q} \partial \mathbf{q}^T} = -\sum_{j=1}^{N} \frac{f(z^j)}{(1 - f(z^j))^2} (1 - e^j) \mathbf{c}^j \mathbf{c}^{j\,T} \leq 0.$$

Hence, the function h is concave, and the log-likelihood for the noisy AND model has no other stationary points than global maxima.

3.4.2 General case

The EM algorithm is guaranteed to converge to the local maxima or saddle points. Thus, we can only be sure that the global maximum, i.e. a point $\boldsymbol{\theta}^*$ such that $CLL(\boldsymbol{\theta}^*) \geq CLL(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}$, will be found if the log-likelihood has neither saddle points nor local maxima. However, the log-likelihood function for a causal independence model with any symmetric Boolean function does not always fulfill this requirement as is shown in the following counterexample.

Example 1 Let us assume a data set $\mathbf{D} = \{(1,1,1,1),(1,0,1,0)\}$ and the interaction function ϵ_1, i.e. $\gamma_1 = 1$ and $\gamma_0 = \gamma_2 = \gamma_3 = 0$. To learn the unknown parameters in the model describing this interaction, we have to maximize the conditional log-likelihood function

$$\begin{aligned} CLL(\boldsymbol{\theta}) &= \ln[\theta_1(1-\theta_2)(1-\theta_3) + (1-\theta_1)\theta_2(1-\theta_3) + (1-\theta_1)(1-\theta_2)\theta_3] \\ &+ \ln[1-\theta_1(1-\theta_3) - (1-\theta_1)\theta_3]. \end{aligned}$$

Depending on the choice for initial parameter settings $\boldsymbol{\theta}^{(0)}$, the EM algorithm converges to one of the maxima:

$$CLL(\boldsymbol{\theta})_{\max} = \begin{cases} 0 & \text{at} \quad \boldsymbol{\theta} = (0,1,0), \\ -1.386 & \text{at} \quad \boldsymbol{\theta} \in \{\left(\theta_1, 0, \frac{1}{2}\right), \left(\frac{1}{2}, 0, \theta_3\right)\}. \end{cases}$$

Obviously, only the point $\boldsymbol{\theta} = (0,1,0)$ is a global maximum of the log-likelihood function while the other obtained points are local maxima.

The existence of local maxima can also be shown graphically. Let us take the two points $\boldsymbol{\theta}' = (0,1,0)$ and $\boldsymbol{\theta}'' = \left(\frac{4}{5}, 0, \frac{1}{2}\right)$, then the resulting log-likelihood given the convex combinations of $\boldsymbol{\theta}'$ and $\boldsymbol{\theta}''$

$$CLL(\lambda\boldsymbol{\theta}' + (1-\lambda)\boldsymbol{\theta}''), \lambda \in [0,1]$$

is convex as shown in Figure 3.1. □

The discussed counterexample proves that in the general case the EM algorithm for symmetric causal independence models does not necessarily converge to a global maximum.

3.5 Experimental results

The EM algorithm developed in this chapter allows us to assess the practical significance of symmetric causal independence models. To do so, we evaluated symmetric causal independence models on the basis of their classification performance.

For our experiments, we chose to use two data sets that are different in their causal interpretation and size. The non-Hodgkin lymphoma data set

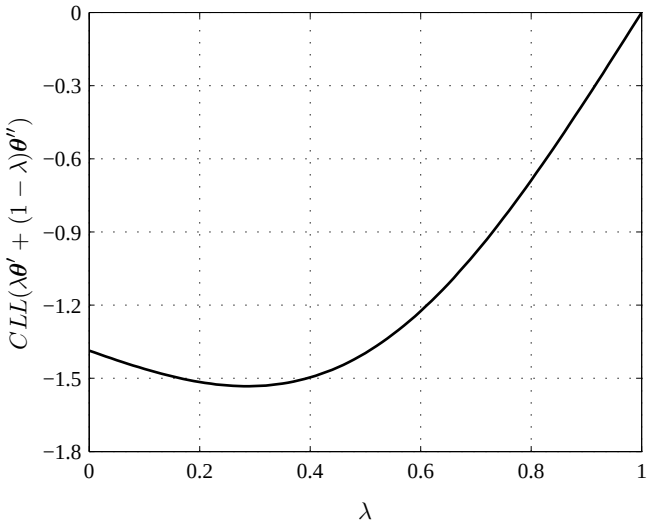


Figure 3.1: The log-likelihood $CLL(\lambda\boldsymbol{\theta}' + (1-\lambda)\boldsymbol{\theta}'')$ from Example 1 as a function of λ.

consists of the factors that influence the result of the treatment, and, for this reason, the models learned from this data set can be argued to follow the causal interpretation. The second data set consisting of Reuters news stories does not follow the causal interpretation. This data set is important because of its size; experiments on the Reuters data collection allowed us to test the EM algorithm on large symmetric causal independence models where the number of cause variables for some document classes is in the hundreds.

3.5.1 Evaluation scheme

We modelled the interaction among cause and effect variables by means of Boolean threshold functions, which seem to be the most probable interaction functions for the given domains. However, the models of document classes in Reuters data set had tens or even hundreds of causes, making learning the models with all threshold functions computationally expensive. Therefore, for this data collection, we only learned the models with the threshold functions τ_2, τ_3, τ_4, the closest threshold functions to the OR function, which was shown to perform well on this data collection [118]. To give a feeling to what extent classification performance of symmetric

causal independence models is influenced by the choice of an interaction function, we report all results obtained.

Given the model parameters $\boldsymbol{\theta}$, the testing data $\mathbf{D}_{test}$ and the classification threshold $\frac{1}{2}$, the classifications and misclassifications for both classes are computed. Let *tp (true positives)* stand for the number of data samples $(\mathbf{c}^j, e^{j+}) \in \mathbf{D}_{test}$ for which $\Pr(e^+ \mid \mathbf{c}^j, \boldsymbol{\theta}) \geq \frac{1}{2}$, and *fn (false negatives)* stand for the number of data samples $(\mathbf{c}^j, e^{j+}) \in \mathbf{D}_{test}$ for which $\Pr(e^+ \mid \mathbf{c}^j, \boldsymbol{\theta}) < \frac{1}{2}$. Likewise, *tn (true negatives)* is the number of data samples $(\mathbf{c}^j, e^{j-}) \in \mathbf{D}_{test}$ for which $\Pr(e^+ \mid \mathbf{c}^j, \boldsymbol{\theta}) < \frac{1}{2}$, and *fp (false positives)* is the number of data samples $(\mathbf{c}^j, e^{j-}) \in \mathbf{D}_{test}$ for which $\Pr(e^+ \mid \mathbf{c}^j, \boldsymbol{\theta}) \geq \frac{1}{2}$. To evaluate the classification performance we used the *accuracy*

$$\eta = \frac{tp + tn}{tp + tn + fn + fp},$$

which is a measure of correctly classified cases, the *F-measure*, which combines *precision* $\pi = \frac{tp}{tp+fp}$ and *recall* $\rho = \frac{tp}{tp+fn}$ as

$$F = \frac{2\pi\rho}{\pi + \rho},$$

and the *area under the ROC curve* (AUC), which is estimated by a generalization of the Mann-Whitney U statistic [3]

$$AUC = \frac{\sum_{(\mathbf{c}^i, e^{i+}) \in \mathbf{D}_{test}} \sum_{(\mathbf{c}^j, e^{j-}) \in \mathbf{D}_{test}} \left(\mathbf{I}_{\{\Pr(e^+|\mathbf{c}^i) > \Pr(e^+|\mathbf{c}^j)\}} + \frac{1}{2} \mathbf{I}_{\{\Pr(e^+|\mathbf{c}^i) = \Pr(e^+|\mathbf{c}^j)\}} \right)}{(tp + fn)(tn + fp)},$$

where $\mathbf{I}_{\{.\}}$ denotes the indicator variable whose value is one if its argument is true and zero otherwise.

To measure the significance of any difference between the classification performance of symmetric causal independence models and other classifiers, we used the exact version of McNemar's test [96]. Let n be the number of cases for which the two classifiers produce different output, and let s be the number of cases where the output of the classifier with higher accuracy was correct, while the output of the other classifier was wrong. Under the null hypothesis that the two classifiers perform equally well, we computed the two-sided p-value

$$p = 2 \sum_{i=s}^{n} \frac{n!}{i!(n-i)!} (0.5)^n.$$

3.5.2 Non-Hodgkin lymphoma data set

This data set contains data from patients with gastric non-Hodgkin lymphoma (NHL), collected by clinical experts from the Netherlands Cancer Institute (see [72] for a thorough description of the disease and collected data).

Gastric non-Hodgkin lymphoma is a type of cancer of the lymphatic system, the disease-fighting network spread throughout the body, which originates in the stomach. Response to treatment is one of the most important prognostic indicators of a long-term disease-free survival, particularly in patients with aggressive NHL [4]. We learned the symmetric causal independence model that models the interaction between the early outcome of the treatment and the pretreatment prognostic factors. The early outcome of the treatment, i.e. the effect in the model, denotes the endoscopically verified result of the treatment, six to eight weeks after treatment; the positive state of this variable, complete remission, defines a situation in which all clinical signs of the disease disappear with the treatment. The following pretreatment prognostic factors are available: (1) age; (2) general health status; (3) bulky disease; (4) histological classification; (5) stage of the cancer; (6) clinical signs (hemorrhage, perforation, obstruction due to the disease). The prognostic factors correspond to the cause variables in the model.

Based on medical literature, we converted the data to binary form and defined each variable such that a false state corresponds to a risk factor that accounts for an impaired complete remission rate. The resulting model is shown in Figure 3.2; the name of the variable indicates a true state of the cause or effect. To learn the parameters of the model we used 125 patient cases with no missing data. 95 of the patients had complete remission six to eight weeks after the treatment, the other 30 patients failed to achieve complete remission. As the data set is small, we could use a leave-one-out cross-validation scheme both to test the performance of the model and to avoid data overfitting. Classification performance measures for symmetric causal independence models with the interaction function $\tau_k,\ k = 1, \ldots, 7$ are reported in Table 3.1. The results indicate that the interaction between the pretreatment variables and the outcome of the treatment is best modelled by the interaction function τ_2. Note that the symmetric causal independence model with the function τ_2 outperforms the noisy OR model, while the noisy AND model is a poor choice to model the given problem.

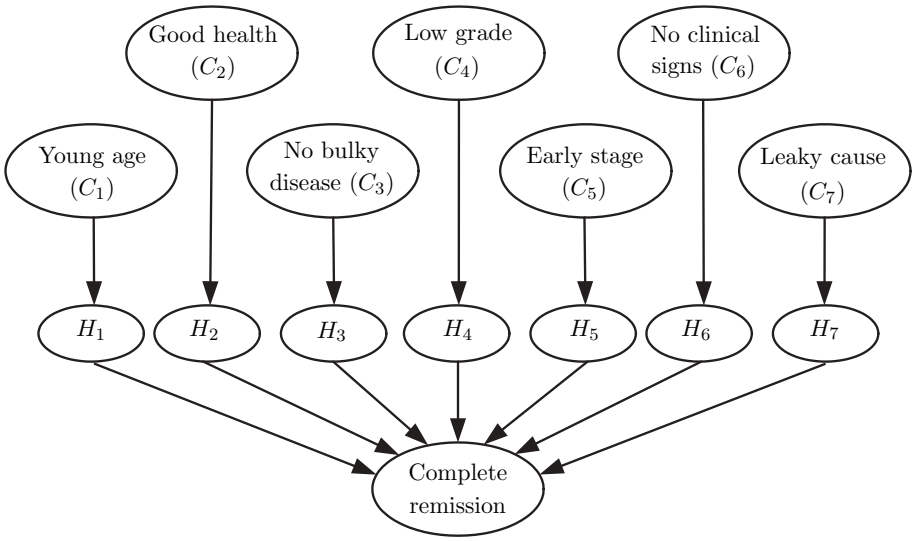


Figure 3.2: Causal independence model modelling complete remission following the treatment of Non-Hodgkin lymphoma. The variable 'Young age' represents a patient younger than 60 years, the variable 'Early stage' stands for the first clinical stage of NHL, and the variable 'No clinical presentation' represents a patient who has no hemorrhage, no perforation and no obstruction.

Table 3.1: Classification performance measures for symmetric causal independence models (SCIMs) with the interaction function $\tau_k, k = 1, \ldots, 7$ for the non-Hodgkin lymphoma data set.

Classifier	Accuracy (%)	F-measure	AUC
SCIM, τ_1 (noisy OR)	75.2	0.854	0.785
SCIM, τ_2	83.2	0.896	0.832
SCIM, τ_3	82.4	0.891	0.834
SCIM, τ_4	78.4	0.857	0.775
SCIM, τ_5	71.2	0.795	0.733
SCIM, τ_6	56.8	0.625	0.661
SCIM, τ_7 (noisy AND)	36.8	0.288	0.584

To compare symmetric causal independence models with other classification algorithms, we evaluated the classification performance of a few widely-used classifiers. The experiments were performed using the Weka system with its default settings [122]. The results reported in Table 3.2 show that the classification performance of the symmetric causal independence model with the function τ_2 is very similar to that of naive Bayes, logistic regression and the multilayer perceptron. However, this symmetric causal independence model outperformed the noisy OR model, decision tree and support vector machine with two-sided p-values of 0.0063, 0.0042 and 0.0923, respectively.

Table 3.2: Classification performance measures for different classifiers for the non-Hodgkin lymphoma data set.

Classifier	Accuracy (%)	F-measure	AUC
SCIM, τ_2	83.2	0.896	0.832
naive Bayes	84.0	0.899	0.821
logistic regression	82.4	0.885	0.823
multilayer perceptron	82.4	0.885	0.780
decision tree (C4.5)	73.6	0.832	0.708
support vector machine	77.6	0.861	0.625

3.5.3 Reuters data set

For the second part of our experiments, we use the Reuters-21578 text categorization collection containing the Reuters news stories preprocessed by Karčiauskas [58]. The comparison of the noisy OR model and a few widely-used classifiers on this data collection was reported in [118], and the results showed the competitive performance of the noisy OR model. Therefore, we aim to show that extended symmetric causal independence models perform competitively with the noisy OR model.

The data set has been split into training (7769 documents) and test (3018 documents) sets. For every one of the ten document classes, the most informative features were selected using the expected information gain as a feature selection criteria. Each document class was classified separately against all other classes. We chose to use the same threshold for the expected information gain as in [118]. The number of selected features varied from 23 for the *corn* document class to 307 for the *earn* document class. Classification performance measures for symmetric causal independence models with the interaction function $\tau_k,\ k = 1, \ldots, 4$ are given in Tables 3.3, 3.4 and 3.5. Even though the threshold to select the relevant features was tuned for the noisy OR model, for five document classes a symmetric causal independence model with another interaction function than the OR function provides better results. For the documents classes *earn* and *trade* the difference in performance is significant with double-sided p-values of 0.0016 (SCIM, τ_2) and 0.0154 (SCIM, τ_4), respectively.

3.6 Discussion

In this chapter, we have developed a computationally efficient EM algorithm to learn parameters in symmetric causal independence models, where the computational scheme of the Poisson binomial distribution was used for the computation of the conditional probabilities in the expectation step. We also investigated the maxima of the log-likelihood function for symmetric causal independence models and showed that the log-likelihood for the noisy OR and the noisy AND models has only global maxima. The presented algorithm allowed us to evaluate the utility of the extended symmetric causal independence models. The reported experimental results indicate that it is unnecessary to restrict causal independence models to only two interaction functions, logical OR and logical AND. Competi-

Table 3.3: Accuracy of symmetric causal independence models (SCIMs) with the interaction function $\tau_k, k = 1, \ldots, 4$ for the Reuters data set; N_{Class} is the number of documents in the corresponding class. The highest accuracy obtained for each document class is shown in bold.

Class	N_{Class}	noisy OR	SCIM, τ_2	SCIM, τ_3	SCIM, τ_4
earn	1087	96.3	**97.2**	**97.2**	96.8
acq	719	93.1	**93.2**	**93.2**	93.0
crude	189	**98.1**	**98.1**	97.6	97.7
money-fx	179	95.8	95.8	95.9	**96.0**
grain	149	**99.2**	99.0	98.2	97.9
interest	131	96.5	**96.8**	96.7	96.7
trade	117	96.6	97.0	**97.3**	**97.3**
ship	89	**98.9**	98.8	98.7	98.6
wheat	71	**99.5**	99.2	98.8	98.5
corn	56	**99.7**	99.4	99.1	98.8

tive performance of the extended symmetric causal independence models present them as a potentially useful additional tool to the set of classifiers.

Even though we described symmetric causal independence models as models constructed on the basis of different behavioural patterns among causes and effects, this description should not limit the use of the framework. When a causal interpretation cannot be applied, symmetric causal independence models can be used as merely a technique to reduce the number of parameters and to simplify the inference problem in Bayesian networks.

The current study has examined the problem of parameter learning in symmetric causal independence models, but the problem of learning the optimal interaction function has not been addressed. Efficient search in the symmetric Boolean function space is a possible direction for further research.

The EM algorithm presented in this chapter learns parameters in models where both cause and effect variables are assumed to be binary. However, causal independence models do not have to be limited to binary variables. Researchers proposed several schemes to generalize the noisy OR model to multivalued variables [27, 45, 105]. Extension of the framework of symmetric causal independence models to handle multivalued variables and

Table 3.4: F-measure of symmetric causal independence models with the interaction function $\tau_k, k = 1, \ldots, 4$ for the Reuters data set; N_{Class} is the number of documents in the corresponding class. The highest F-measure obtained for each document class is shown in bold.

Class	N_{Class}	noisy OR	SCIM, τ_2	SCIM, τ_3	SCIM, τ_4
earn	1087	95.0	**96.1**	**96.1**	95.6
acq	719	**85.3**	84.3	84.5	83.8
crude	189	84.5	**85.7**	80.7	81.0
money-fx	179	60.9	62.1	62.6	**62.7**
grain	149	**92.7**	89.9	80.7	77.2
interest	131	40.2	**55.0**	53.3	54.0
trade	117	51.0	61.2	**63.7**	**63.7**
ship	89	**79.5**	77.7	74.5	71.5
wheat	71	**90.3**	81.8	71.4	66.2
corn	56	**91.8**	83.6	72.5	61.5

adjustment of the proposed EM algorithm to this generalization is another research problem of interest.

Acknowledgments

We are grateful to Henk Boot and Babs Taal for providing the non-Hodgkin's lymphoma data, Gytis Karčiauskas for the preprocessed Reuters data and Jiří Vomlel for sharing his code and insights.

3.7 Appendix

Even if we use recursive methods to compute the Poisson binomial distribution, the computation of the probabilities $\Pr(h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)}), k = 1, \ldots, n$ for a data sample $\mathbf{x}^j$ requires $O(n^3)$ operations. This appendix explains how the theory of the Poisson binomial distribution can be applied to further reduce the computational complexity of the expectation step of the EM algorithm for symmetric causal independence models.

Table 3.5: Area under the ROC curve of symmetric causal independence models with the interaction function $\tau_k, k = 1, \ldots, 4$ for the Reuters data set; N_{Class} is the number of documents in the corresponding class. The highest AUC obtained for each document class is shown in bold.

Class	N_{Class}	noisy OR	SCIM, τ_2	SCIM, τ_3	SCIM, τ_4
earn	1087	0.995	**0.996**	0.995	0.992
acq	719	**0.972**	**0.972**	0.957	0.928
crude	189	0.994	**0.995**	0.994	0.983
money-fx	179	0.971	**0.973**	0.957	0.915
grain	149	**0.999**	0.997	0.985	0.952
interest	131	0.961	**0.963**	0.949	0.915
trade	117	0.979	**0.985**	0.982	0.973
ship	89	0.979	**0.986**	0.938	0.842
wheat	71	**0.997**	0.995	0.994	0.956
corn	56	**0.998**	0.996	0.982	0.939

3.7.1 Reducing the size of the input

The obvious way to reduce the size of the input is to remove those probabilistic parameters from $\mathbf{p}$ that equal zero or one and adapt the Boolean constants accordingly. Let us define three new vectors:

$$\begin{aligned}
\mathbf{p}_{\setminus(0)} &= (p_i; i = 1 \ldots n, p_i \neq 0), \\
\mathbf{p}_{\setminus(1)} &= (p_i; i = 1 \ldots n, p_i \neq 1), \\
\mathbf{p}_{\setminus(0,1)} &= (p_i; i = 1 \ldots n, p_i \neq 0, p_i \neq 1).
\end{aligned}$$

Using Equation (3.14) iteratively, we obtain the relationships $\mathrm{B}(i; \mathbf{p}) = \mathrm{B}(i - n_1; \mathbf{p}_{\setminus(1)})$ and $\mathrm{B}(i; \mathbf{p}) = \mathrm{B}(i; \mathbf{p}_{\setminus(0)})$, where n_1 is the number of elements of $\mathbf{p}$ equal to 1. Combining these two results yields:

$$\mathrm{B}(i; \mathbf{p}) = \mathrm{B}(i - n_1; \mathbf{p}_{\setminus(0,1)}).$$

Since $\mathrm{B}(i - n_1; \mathbf{p}_{\setminus(0,1)}) = 0$ for all $i < n_1$ and $i > n - n_0$ (with n_0 being the number of zero elements of $\mathbf{p}$), we can write the conditional probability of

the effect as

$$\Pr(e^+ \mid \mathbf{c}^j, \boldsymbol{\theta}^{(z)}) = \sum_{i=0}^{n} \mathrm{B}(i - n_1; \mathbf{p}_{\setminus(0,1)}^{(z,j)}) \gamma_i = \sum_{i=0}^{n-n_0-n_1} \mathrm{B}(i; \mathbf{p}_{\setminus(0,1)}^{(z,j)}) \gamma_{i+n_1}. \quad (3.19)$$

Equation (3.19) allows the reduction of the size of the input from n to $n - n_0 - n_1$. Note that, given the assumption $\Pr(h_i^+ \mid c_i^-) = 0$, n_0 is equal to or larger than the number of 'inactive' causes in a given data sample $\mathbf{x}^j$.

3.7.2 Specific cases

From Equation (3.19) and the Poisson binomial identity (3.14), we derived a few special cases where we do not need to compute the Poisson binomial distribution in order to compute the probability $\Pr(h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)})$:

$$\Pr(h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)}) = \begin{cases} p_k^{(z,j)} & \text{if } p_k^{(z,j)} \in \{0, 1\}, \\ \frac{p_k^{(z,j)}}{1 - p_k^{(z,j)}} & \text{if } \max(u \mid \gamma_u = e^j, u = 0, \ldots, n - n_0) = 0, \\ 1 & \text{if } \min(u \mid \gamma_u = e^j, u = 0, \ldots, n - n_0) = n - n_0. \end{cases} \quad (3.20)$$

3.7.3 Number of operations

At the expectation step of the $(z+1)$-th iteration of the EM algorithm for a given data sample $\mathbf{x}^j$, we need to compute the Poisson binomial distributions

$$\mathrm{B}(0; \mathbf{p}^{(z,j)}), \ldots, \mathrm{B}(n; \mathbf{p}^{(z,j)}), \text{ and}$$

$$\mathrm{B}(0; \mathbf{p}_{\setminus k}^{(z,j)}), \ldots, \mathrm{B}(n-1; \mathbf{p}_{\setminus k}^{(z,j)}).$$

Given (3.20), we need to compute only those distributions $\mathrm{B}(0; \mathbf{p}_{\setminus k}^{(z,j)}), \ldots, \mathrm{B}(n-1; \mathbf{p}_{\setminus k}^{(z,j)})$ for which $p_k^{(z,j)} \notin \{0, 1\}$. Using the recursive method to compute the Poisson binomial distribution and Equation (3.19), computation of each Poisson binomial distribution requires $(n - n_0 - n_1)^2$ operations. The data sample $\mathbf{x}^j$ requires at most $(n - n_0 - n_1)^3$ operations. Thus,

the computational complexity of the EM algorithm was reduced but it remained cubic.

The number of operations can be reduced from cubic to quadratic by the use of Equation (3.14). To do so, we first need to compute the distribution $\mathrm{B}(0; \mathbf{p}^{(z,j)}), \ldots, \mathrm{B}(n; \mathbf{p}^{(z,j)})$, and then iteratively calculate $\mathrm{B}(0; \mathbf{p}^{(z,j)}_{\setminus k}), \ldots,$ $\mathrm{B}(n-1; \mathbf{p}^{(z,j)}_{\setminus k})$ either from

$$\mathrm{B}(i; \mathbf{p}^{(z,j)}_{\setminus k}) = \frac{\mathrm{B}(i; \mathbf{p}^{(z,j)}) - \mathrm{B}(i-1; \mathbf{p}^{(z,j)}_{\setminus k}) p_k^{(z,j)}}{1 - p_k^{(z,j)}} \tag{3.21}$$

starting with $\mathrm{B}(-1; \mathbf{p}^{(z,j)}_{\setminus k}) = 0$, or from

$$\mathrm{B}(i-1; \mathbf{p}^{(z,j)}_{\setminus k}) = \frac{\mathrm{B}(i; \mathbf{p}^{(z,j)}) - \mathrm{B}(i; \mathbf{p}^{(z,j)}_{\setminus k})(1 - p_k^{(z,j)})}{p_k^{(z,j)}} \tag{3.22}$$

starting with $\mathrm{B}(n; \mathbf{p}^{(z,j)}_{\setminus k}) = 0$.

This scheme works because Equation (3.14) reduces to $\mathrm{B}(0; \mathbf{p}^{(z,j)}) = \mathrm{B}(0; \mathbf{p}^{(z,j)}_{\setminus k})(1-p_k^{(z,j)})$ and $\mathrm{B}(n; \mathbf{p}^{(z,j)}) = \mathrm{B}(n-1; \mathbf{p}^{(z,j)}_{\setminus k}) p_k^{(z,j)}$ when $i = 0$ and $i = n$, respectively. Note that we know $p_k^{(z,j)}$ and that we have previously computed $\mathrm{B}(0; \mathbf{p}^{(z,j)}), \ldots, \mathrm{B}(n; \mathbf{p}^{(z,j)})$. To reduce the effect of roundoff error, which may occur using the iterative method, we advise to use the bottom-up approach (3.21) when $p_k^{(z,j)} < \frac{1}{2}$ and the top-down approach (3.22) when $p_k^{(z,j)} > \frac{1}{2}$.

Using the results presented in this appendix, the probability $\Pr(h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)})$ can be computed in at most $O((n - n_0 - n_1)^2)$ operations.

Chapter 4

Modelling Carcinoid Heart Disease

Carcinoid heart disease is the most dangerous complication of carcinoid syndrome as it occurs in over 65% of patients with carcinoid syndrome and is a major source of morbidity and mortality for these patients. We use noisy threshold models to predict the development of carcinoid heart disease. We use data of fifty-four patients who suffered from a low-grade midgut carcinoid tumor, of which twenty-two patients developed carcinoid heart disease. Eleven attributes that are known at admission have been used to classify whether the patient develops carcinoid heart disease.
The noisy threshold model performed favorably to other state-of-the-art classification algorithms and equally well as a decision rule that was formulated by the physician.

This chapter is based on: M.A.J. van Gerven, R. Jurgelenaite, B.G. Taal, T. Heskes and P.J.F. Lucas. Predicting carcinoid heart disease with the noisy threshold classifier. *Artificial Intelligence in Medicine*, 2007.

4.1 Introduction

Bayesian networks have become a widely accepted formalism for reasoning under uncertainty by providing a concise representation of a joint proba-

bility distribution over a set of random variables [85]. This distribution is factorized according to an associated acyclic directed graph (ADG) that represents the independence structure between random variables. However, the construction of a Bayesian network that fully captures this independence structure for a realistic domain, has proven to be a difficult task. It requires either manual specification of the ADG by means of available expert knowledge, or large amounts of high-quality data when we resort to structure learning.

An alternative to the construction of an ADG that fully captures the independence structure that holds between variables within the domain, is to use a fixed or severely constrained graph topology for classification purposes. In the latter context we call a Bayesian network a Bayesian classifier. The use of Bayesian methods in medicine was first proposed by Ledley and Lusted in their classic 1959 paper [70], and one of the first successful implementations of Bayesian classifiers in medicine was De Dombals system for the diagnosis of acute abdominal pain [23]. The classifier that was used assumes independence of symptoms given the disease, and is known as the naive Bayes classifier. Over the years, many different Bayesian classifier architectures have been proposed [16, 34, 94, 104].

Although, typically, the actual joint probability distribution, and the joint probability distribution that is represented by the Bayesian classifier, differ considerably, this approach can still yield good results with respect to the classification task [108]. However, a weakness of this approach is that the ad-hoc restrictions that are placed on the underlying graph effectively reduces the Bayesian network to a black box model, making the relation between properties of the domain and classification outcome often difficult to understand. This is an undesirable property; especially in medicine, where ideally one wants to be able to interpret how the classification outcome (such as diagnosed disease or patient prognosis) relates to the available domain knowledge (its causes). The explanation of drawn conclusions is required to increase the acceptance of machine-learning techniques in practice [65].

In this chapter, we employ a noisy threshold model for classification tasks. This noisy threshold classifier is based on a generalization of the well-known noisy OR model, which has already been used for the purpose of text classification in [117]. In order to demonstrate the merits of the noisy threshold classifier in a medical context, we apply the technique to the prediction of carcinoid heart disease (CHD); a serious condition that arises as a complication of certain neuroendocrine tumors [130]. We demonstrate that the

noisy threshold classifier performs competitively with state-of-the-art classification techniques for this medically relevant problem. Furthermore, an expert physician at the Netherlands Cancer Institute (NKI) was consulted, and it is demonstrated how her knowledge concerning CHD relates to the parameters that were estimated for the noisy threshold classifier.

This chapter proceeds as follows. Section 4.2 describes the medical problem. The use of the noisy threshold model as a Bayesian classifier is discussed in Section 4.3. The results on the classification task and the medical interpretation by the expert physician is presented in Section 4.4. The chapter is ended by some concluding remarks in Section 4.5.

4.2 Carcinoid heart disease

Carcinoid tumors belong to the group of neuroendocrine tumors, which are known for the production of vasoactive agents in the presence of metastatic disease; usually hepatic (liver) metastases. Among these agents, serotonin is the most important agent, leading to the characteristic carcinoid syndrome of flushes and diarrhea. The other main characteristic feature of neuroendocrine tumors is the slow progression of most tumors if the histology shows a low-grade pattern [129].

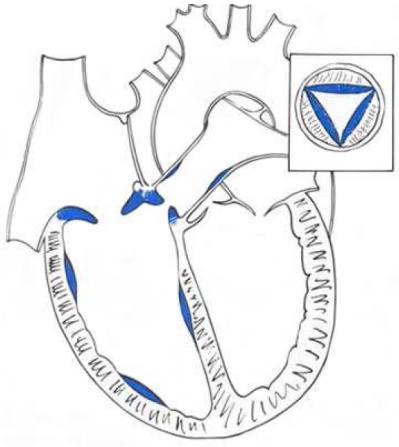


Figure 4.1: CHD is characterized by heart valve fibrosis.

Serotonin overproduction may also cause carcinoid heart disease (CHD), which is characterized by fibrosis of the right sided heart valves as shown in Figure 4.1. Fibrosis induces thickening and retraction of the tricuspid

valve, leading to tricuspid insufficiency and ultimately heart failure, which is the cause of death in as much as half of carcinoid patients [129, 130]. Since so many carcinoid patients die of CHD, it is important to divide patients that are admitted to the clinic into patients that are prone to develop a severe form of carcinoid heart disease, and those that are not expected to develop this severe form. In this way, patients that are at risk can be given more aggressive treatment in order to reduce the probability of the development of CHD. Hence, the classification task for this medical problem will be to classify the patients into these two groups, depending on the attributes that are known at the time of admission to the clinic. We use chd^+ to denote the development of moderate to extreme tricuspid valve insufficiency and chd^- to denote the absence, or development of mild tricuspid valve insufficiency during patient follow-up.

Table 4.1: Patient attributes that are measured at admission.

Name	Definition	Name	Definition
HIA	5-HIAA levels	GIL	General illness
CGA	Chromogranin A levels	BOB	Bowel obstruction
DIA	Diarrhea	IBL	Internal bleeding
WHE	Wheezing	FEV	Fever
FLU	Flushing	HME	Hepatic metastases
APA	Abdominal pain		

In principle, the physician can make use of the attributes that are measured at admission (Table 4.1), in order to predict the development of CHD. However, in practice, in order to determine the probability of developing moderate to severe tricuspid valve insufficiency, the physician makes use of the following decision rule:

$$\Pr(\text{chd}^+ \mid \mathbf{c}) = \begin{cases} 0.50 & \text{if } \text{hia}^+ \wedge \text{dia}^+ \wedge \text{hme}^+ \\ 0.25 & \text{if } \text{hia}^+ \wedge (\text{dia}^- \wedge \text{hme}^+ \vee \text{dia}^+ \wedge \text{hme}^-) \\ 0.10 & \text{if } \text{hia}^+ \wedge \text{dia}^- \wedge \text{hme}^- \vee \text{hia}^- \wedge \text{dia}^+ \wedge \text{hme}^+ \\ 0.03 & \text{otherwise.} \end{cases}$$

The aim of this chapter is to show that a noisy threshold model can be used as a Bayesian classifier, where performance is compared both with

the physicians classification performance, as well as with standard classification techniques such as the naive Bayes classifier, logistic regression and decision trees. The patient attributes are used as cause variables in the definition of a noisy threshold model. As required, variables are binary, and positive states of variables are perceived to be less favorable than negative states, such that they could be responsible for carcinoid heart disease. To train and test Bayesian classifiers for this medical problem, we have used a clinical database consisting of fifty-four patients that suffered from a neuroendocrine tumor, and for which the grade of tricuspid valve insufficiency was known. Twenty-two patients developed moderate or worse tricuspid valve insufficiency during follow-up.

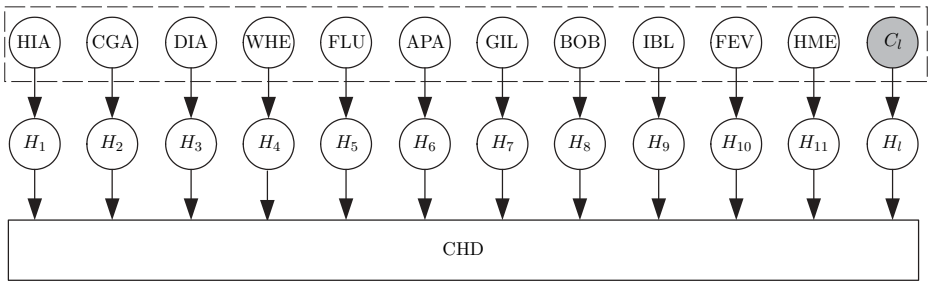


Figure 4.2: A noisy threshold model for carcinoid heart disease, where the dashed region represents the total tumor burden for the patient. Note the use of the leak cause C_l in order to model possible hidden causes.

We have not yet touched upon the most important assumption of causal independence models. That is, can the variables be regarded as causes of carcinoid heart disease? For some attributes this is questionable. Diarrhea for instance is a symptom of other processes and is therefore not likely to be a cause of carcinoid heart disease. However, we can interpret the attributes as risk factors that act as components of the total tumor burden, as depicted in Figure 4.2. Since the causes are assumed to be completely observed, we refrain from adding additional dependencies between cause variables.

4.3 The noisy threshold classifier

4.3.1 Classifier construction

Construction of a noisy threshold classifier (NTC) is as follows. We first determine the cause variables $\mathbf{C}$ and effect variable E that are used in the classifier. In the context of a classifier, the cause variables stand for the attributes and the effect variable stands for the class-variable. Secondly, we need to determine the positive states of the variables. In the CHD domain, the positive states are simply defined as the presence of attributes that affect the presence of the class-variable CHD. Once the cause and effect variables have been defined, we need to find both the optimal values for the parameters $\Pr(h_i^+ \mid c_i^+)$ using the EM algorithm presented in Chapter 3, as well as the best suited threshold function τ_k.

Recall that the $(z+1)$-th iteration of the EM algorithm for symmetric causal independence models is given by:
Expectation step: For every instance $\mathbf{x}^j = (\mathbf{c}^j, e^j)$ with $j = 1, \ldots, N$,

we form

$$\mathbf{p}^{(z,j)} = (p_1^{(z,j)}, \ldots, p_n^{(z,j)}) \quad \text{where} \quad p_i^{(z,j)} = \theta_i^{(z)} c_i^j. \tag{4.1}$$

Subsequently, the probability $\mathrm{P}(h_k^+ \mid \mathbf{c}^j, e^j, \boldsymbol{\theta}^{(z)})$ is computed for each hidden variable $H_k, k = 1, \ldots, n$ from

$$\Pr(h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)}) = \begin{cases} \dfrac{p_k^{(z,j)} \sum_{i=0}^{n-1} \mathrm{B}\left(i; \mathbf{p}_{\setminus k}^{(z,j)}\right) \gamma_{i+1}}{\sum_{i=0}^{n} \mathrm{B}\left(i; \mathbf{p}^{(z,j)}\right) \gamma_i} & \text{if } e^j = 1, \\ \dfrac{p_k^{(z,j)} \left(1 - \sum_{i=0}^{n-1} \mathrm{B}\left(i; \mathbf{p}_{\setminus k}^{(z,j)}\right) \gamma_{i+1}\right)}{1 - \sum_{i=0}^{n} \mathrm{B}\left(i; \mathbf{p}^{(z,j)}\right) \gamma_i} & \text{if } e^j = 0. \end{cases} \tag{4.2}$$

Maximization step: Update the parameter estimates using

$$\theta_k = \frac{\sum_{1 \le j \le N} \sum_{\mathbf{h}_{\setminus k}} \Pr(\mathbf{h}_{\setminus k}, h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)})}{\sum_{1 \le j \le N} c_k^j \sum_{\mathbf{h}_{\setminus k}} \left(\Pr(\mathbf{h}_{\setminus k}, h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)}) + \Pr(\mathbf{h}_{\setminus k}, h_k^- \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)})\right)}$$

$$= \frac{\sum_{1\leq j\leq N} \Pr(h_k^+ \mid e^j, \mathbf{c}^j, \boldsymbol{\theta}^{(z)})}{\sum_{1\leq j\leq N} c_k^j}. \tag{4.3}$$

To evaluate the classification performance of the classifiers, we use three measures, accuracy η, F-measure F and AUC as given in Chapter 3, since the accuracy is the obvious measure but may convey the wrong intuition when the classes are not equal in size [115]. Finding the optimal noisy threshold classifier then proceeds as follows:

1. Divide the data set $\mathbf{D}$ into the disjoint sets $\mathbf{D}_{train}$, $\mathbf{D}_{validate}$ and $\mathbf{D}_{test}$.
2. For all noisy threshold models $k = 1, \ldots, n$ use the training data $\mathbf{D}_{train}$ and the EM algorithm of Chapter 3 to learn the parameters $\Pr(h_i^+ \mid c_i^+)$.
3. Select the noisy threshold model and the number of iterations of the EM algorithm that maximizes $w_1\eta(\mathbf{D}_{validate}) + w_2F(\mathbf{D}_{validate})$ with equal weights $w_1 = w_2$, as the optimal noisy threshold classifier.

With regard to the clinical data set $\mathbf{D}$ we have used a leave-one-out cross-validation scheme to implement the above algorithm. $\mathbf{D}$ contains too many missing values to simply remove the instances that contain missing data. We have used *mean substitution* [62] as an imputation scheme, and note that *multiple imputation* [91] produced similar results. Let N_i be the number of data samples without missing data for the variable C_i for all $i = 1, \ldots, n$. If the value for the variable C_i is missing in the sample j, then we replace c_i^j in Equations (4.1) and (4.3) by the estimate

$$\mu_i = \frac{1}{N_i} \sum_{k=1}^{N_i} c_i^k$$

of the prior $\Pr(c_i^+)$.

4.3.2 Classifier evaluation

In order to evaluate the performance of the noisy threshold classifier, we compare its classification accuracy with the accuracy of a number of other

well-known algorithms. For the comparison we use the naive Bayes classifier (NBC), logistic regression (LG) and a decision tree learning algorithm (C4.5) as implemented by the WEKA machine learning tool [122]. We use WEKAs default parameter settings; the default imputation method is to interpret a missing value for a variable as a separate value. Furthermore, we compare the performance of the optimal noisy threshold classifier with that of the noisy OR classifier, which is a special case of noisy threshold classifiers [117].

To measure the significance of any difference between the classification performance of symmetric causal independence models and other classifiers, we used the exact version of McNemar's test [96] (see Chapter 3).

4.4 Results

4.4.1 Classification performance

Table 4.2 lists the classification accuracy for noisy threshold classifiers with threshold $k = 1, \ldots, n$. The noisy threshold classifier $k = 6$ is selected, based on the validation set $\mathbf{D}_{validate}$, and shows the best classification accuracy of 0.72 on the test set $\mathbf{D}_{test}$. Note that the accuracy is considerably better than the classification accuracy of 0.54 for the noisy OR classifier.

Table 4.2: Classification accuracy on the test set for noisy threshold classifiers.

Noisy threshold classifier	Accuracy (%)	Noisy threshold classifier	Accuracy (%)
τ_1	0.54	τ_7	0.65
τ_2	0.65	τ_8	0.57
τ_3	0.65	τ_9	0.59
τ_4	0.70	τ_{10}	0.59
τ_5	0.69	τ_{11}	0.59
τ_6	0.72	τ_{12}	0.59

In order to test how well the noisy threshold classifier performs compared with the physician, and with other classification algorithms, we have determined the classification accuracy. Table 4.3 describes the classification

accuracy on $\mathbf{D}_{test}$ for the physician, naive Bayes classifier (NBC), logistic regression (LG), decision tree learning algorithm (C4.5) and noisy OR, and p-values for the null-hypothesis that the classifier accuracy is comparable to that of the noisy threshold classifier with threshold $k = 6$.

Table 4.3: Classification accuracy on the test set and p-values for the null hypothesis that a classifier is just as good as the noisy threshold classifier with threshold $k = 6$.

Classifier	Accuracy (%)	p-value
physician	0.69	0.70
NBC	0.63	0.23
LG	0.67	0.63
C4.5	0.44	$6.2 \cdot 10^{-5}$
noisy OR	0.54	$6.4 \cdot 10^{-3}$

Note that the expert physicians classification accuracy is reasonably high, outperforming all but the noisy threshold classifier. The noisy threshold classifier with $k = 6$ shows the best classification accuracy, although the difference is significant only for C4.5 and the noisy OR classifier at a confidence level of $p = 0.05$. For the physicians decision rule, the naive Bayes classifier, and logistic regression, we cannot reject the null hypothesis that the algorithms may in fact be equally accurate for this data set.

It is well-known that classifiers that show large bias tend to outperform classifiers that show high variance for small data sets, since this reduces the risk of overfitting. For this reason, the naive Bayes classifier tends to perform well on many data sets [63]. However, although not always reflected in its classification accuracy [29], the assumption of independence between attributes given the class-variable, is a strong assumption which does not hold in general. In contrast, the noisy threshold classifiers assumptions are motivated by a cause-effect semantics, and hold for domains where the presence of a sufficient number of causes is sufficient to induce the effect.

Figure 4.3 presents the ROC curves for the physician, the noisy threshold classifier with $k = 6$, the naive Bayes classifier and logistic regression, where the area under the curve equals 0.66, 0.66, 0.60 and 0.59 respectively. Although the performance in terms of AUC is mediocre, both the physicians decision rule, and the noisy threshold classifier show a considerably better performance than the other standard classification techniques.

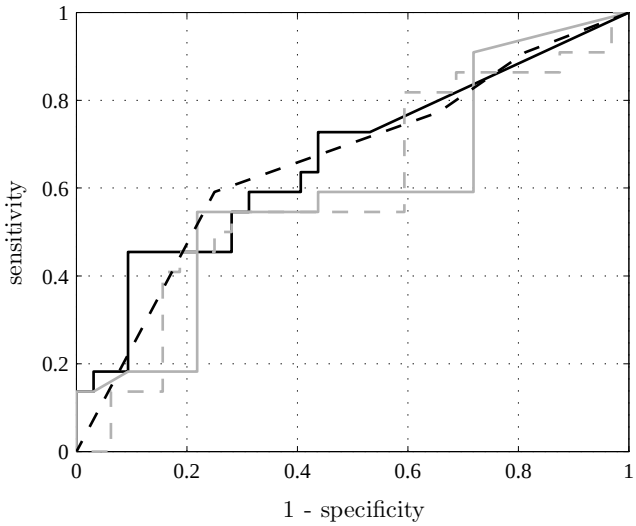


Figure 4.3: ROC curve for the noisy threshold classifier (solid black), logistic regression (solid grey), the naive Bayes classifier (dashed grey) and physician (dashed black) where the straight line segment in the noisy threshold classifier curve is a consequence of the model assumption that absent causes do not contribute to the effect.

The ROC curve does demonstrate a potential danger of using the noisy threshold classifier, especially when the causal assumptions are not satisfied. Whereas the naive Bayes classifier is able to gradually increase the true positive rate at the expense of increasing the true negative rate, the noisy threshold classifier fails to accomplish this for all true positive rates. This is a consequence of the model assumption that absent causes cannot contribute to the effect; the probability $\Pr(e^+ \mid \mathbf{c}^i)$ of assigning an instance to the positive class equals zero whenever the number of present causes is less than the chosen threshold k.

4.4.2 Medical interpretation

In this section, we look at the noisy threshold classifier for CHD from a medical point of view. Prior to presenting the resulting classifier, we have asked the physician to indicate how important the individual attributes were felt to be with respect to predicting the development of carcinoid heart disease.

According to the physician, progressive carcinoid disease is often accompanied by the carcinoid syndrome, which is characterized by diarrhea (DIA) caused by increased bowel motility due to serotonin overproduction, by periodical flushing attacks (FLU) due to the synergistic interaction between various vasoactive agents, and sometimes by wheezing (WHE). As discussed in Section 4.2, serotonin overproduction is thought to play a key role in the etiology of CHD and it can be measured indirectly by means of the urinary 5-HIAA level (HIA) since this is a metabolite of serotonin. Hence, the variables related to the carcinoid syndrome are indicative of serotonin overproduction and ultimately CHD. It is therefore assumed that the variables HIA, DIA, FLU and to a lesser extent WHE have a high predictive value. Serotonin overproduction is itself caused by the carcinoid tumor in the presence of particular metastases; hormones released by carcinoid tumors are often destroyed by the liver before they reach the general circulation to cause symptoms. Therefore, only hepatic metastases (HME), or metastases that can release hormones directly into the general circulation, can produce the carcinoid syndrome. According to the physician, the presence of hepatic metastases (HME) during hospitalization is indicative of CHD development, since this is a requirement for serotonin overproduction. The plasma chromogranin A (CGA) level is used as a general marker of neuroendocrine activity and tumor extensiveness [82]. Although not regarded as important as the previously discussed attributes, the physician expected CGA to have a high predictive value since extensive tumors with high neuroendocrine activity are more likely to cause CHD. In contrast, the variables IBL, FEV, APA and BOB were not thought to predict CHD very well. Local progression of hyper-vascular primary tumors into the lumen of the small bowel is often the cause of internal bleeding (IBL), but is not thought to be related to metastatic disease. Fever (FEV) can be caused by hepatic metastases, as captured by the variable HME, but it is also a non-specific symptom that is not necessarily caused by carcinoid disease in the first place. Abdominal pain (APA) and bowel obstruction (BOB) are often caused by complications due to the primary tumor and were assumed to be unrelated to the development of CHD. According to the physician, general illness (GIL) could be indicative of the development of carcinoid heart disease; a poor condition is often due to extensive metastases and therefore a high probability of serotonin overproduction. In general, the physician expected that at least some of the risk factors should occur together in order to cause CHD.

Figure 4.4 depicts the actual estimates of prior probabilities $\Pr(c_i^+)$ and conditional probabilities $\Pr(h_i^+ \mid c_i^+)$, for the noisy threshold classifier

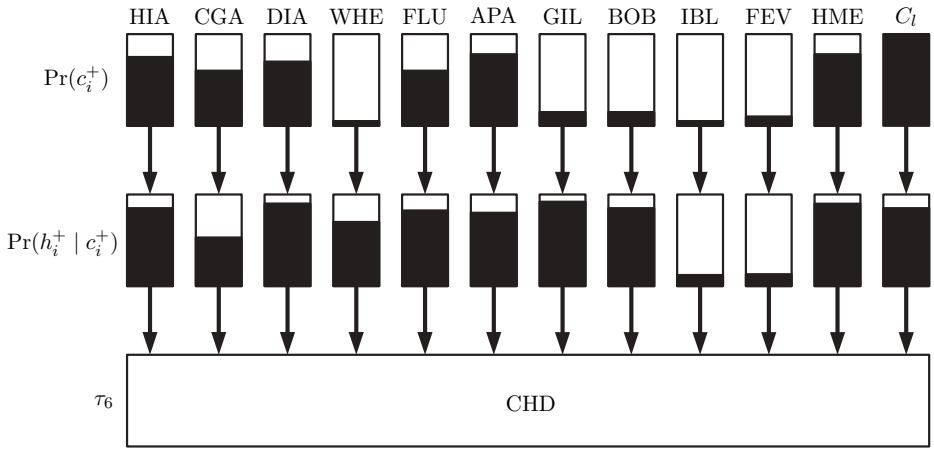


Figure 4.4: Estimates of priors $\Pr(c_i^+)$, and conditional probabilities $\Pr(h_i^+ \mid c_i^+)$, for the noisy threshold classifier with threshold function $k = 6$.

that was used for predicting CHD. The predictive value of the variables HIA, DIA, FLU and WHE is reflected in the reasonably high associated probabilities $\Pr(h_i^+ \mid c_i^+)$ with $i \in \{1, 3, 4, 5\}$, which range from 0.67 to 0.91, where wheezing is indeed seen to be of less predictive value than the other attributes. The presence of hepatic metastases (HME) is also an important predictor of CHD, as is indicated by the high probability $\Pr(h_{11}^+ \mid c_{11}^+) = 0.92$. Notice that most patients that are admitted already present with such metastases, which is reflected by the high prior probability $\Pr(c_{11}^+) = 0.78$. Contrary to the physicians expectations, CGA was not a very good predictor of CHD, with $\Pr(h_i^+ \mid c_i^+) = 0.53$. In hindsight, this may be explained by the fact that CGA overproduction does not necessarily reflect serotonin overproduction, and if it does, it may be redundant information given that we have observed HIA, which is a metabolite of serotonin. Internal bleeding (IBL) and fever (FEV), with $\Pr(h_i^+ \mid c_i^+) = 0.12$ and $\Pr(h_i^+ \mid c_i^+) = 0.13$ respectively did not contribute much to the effect. Unexpectedly, both abdominal pain (APA) and bowel obstruction (BOB) had relatively high probability values $\Pr(h_i^+ \mid c_i^+)$ of 0.80 and 0.84 respectively. After some deliberation, the physician gave the following possible explanation. Since abdominal pain and bowel obstruction are often caused by complications due to the primary tumor, both APA and BOB indicate a midgut tumor with possible mesenterial fibrosis. A midgut localization is a prerequisite for serotonin overproduction, and mesenterial fibrosis is thought to be related to tricuspid valve fibrosis [78], and, therefore, the presence of these variables could have been indicative

of the development of CHD. General illness (GIL) had a high probability value of $\Pr(h_i^+ \mid c_i^+) = 0.93$. Five out of seven patients that suffered from general illness indeed developed CHD. The threshold function $k = 6$ corresponds to the physicians assessment that the presence of just one risk factor is generally insufficient to cause CHD, whereas the presence of all risk factors is much too strict a requirement as a cause for CHD; demonstrating that the noisy threshold model as a generalization of both the noisy OR and noisy AND model can be the proper choice for realistic domains.

4.5 Conclusions

The noisy threshold classifier is a classifier that has a well-defined semantics in terms of causes and effect. Due to the independence assumptions that are made by the classifier, parameters can be reliably estimated without needing to resort to huge amounts of data. This is an important feature since many domains are characterized by limited amounts of data, as discussed in [112]. Learning Bayesian classifiers from data is to be contrasted with the construction of a full Bayesian network that captures available domain knowledge, which, although possible, can be very resource intensive for realistic domains.

We have demonstrated that the noisy threshold classifier performs comparably with the decision rule that is used by an expert physician, and competitively with state-of-the-art classifiers, on an important classification task in oncology. Furthermore, it significantly outperforms the noisy OR classifier, as a special case of the noisy threshold classifier, for this data set. The semantics of the noisy threshold classifier enables an interpretation in terms of available domain knowledge, as is illustrated by the physicians interpretation of classifier parameters. Nevertheless, one should be cautious when defining the positive states of the cause variables since negative states cannot contribute to the effect, as reflected by the straight line segment of the ROC curve. The competitive classification performance and well-defined semantics make the noisy threshold classifier a promising new machine learning technique, as was demonstrated here in the context of medical prognosis.

Chapter 5

Modelling Gene Regulation in Plasmodium Falciparum

To date, there is little knowledge about one of the processes fundamental to the biology of Plasmodium falciparum*, gene regulation including transcriptional control. We use noisy threshold models to identify regulatory sequence elements explaining membership to a gene expression cluster, where each cluster consists of genes active during part of the developmental cycle inside a red blood cell. Differently from other bioinformatics approaches applied to* P. falciparum*, our method is able to model the logic behind gene regulation and to incorporate uncertainty about the functionality of putative regulatory sequence elements. Using the classification accuracy of the noisy threshold models as a measure of their soundness, we test a number of different models and, consequently, different hypotheses about gene regulation. We obtain several interesting results that deserve further (biological) investigation.*

Parts of this chapter appeared in: R. Jurgelenaite, T. Heskes, and T. Dijkstra. Using symmetric causal independence models to predict gene expression from sequence data. In *Proceedings of the ECML-PKDD Workshop 'Data Mining in Functional Genomics and Proteomics: Current Trends and Future Directions'*, 2007.

5.1 Introduction

Malaria is caused by protozoan parasites of which *Plasmodium falciparum* causes up to 2 million deaths, mainly children under the age of 5, annually in sub-Saharan Africa [11, 93]. Malaria is a poverty related disease and vaccines are not yet available [109]. In addition, resistance to the most commonly available antimalarials (chloroquine and antifolate drugs) has spread worldwide. A thorough understanding of the complex biology of the parasite, with developmental stages in humans and Anopheles mosquitoes, will help in designing more effective control strategies to combat malaria. Basic knowledge of one of the processes fundamental to *Plasmodium* biology, gene regulation including transcriptional control, is still lacking. The published *P. falciparum* genome sequence has not provided much grip on transcription control, since not many transcription factors could be identified and intergenic regions appeared to mainly consist of A+T sequences [36].

In this chapter, we present a bioinformatics approach which combines RNA expression data with genome sequence data to identify regulatory sequence elements and to learn more about gene regulatory mechanisms. A key feature of transcriptional regulation of gene expression in eukaryotes (organisms whose cells contain a distinct membrane-bound nucleus) is that genes are often regulated by more than one transcription factor [119]. It has been suggested that combinatorial gene regulation is the general mode of transcriptional regulation in *P. falciparum* [114]. A number of approaches have been proposed to address the combinatorial nature of transcriptional regulation. One group of approaches is based on the assumption that the influence of different transcription factors on gene expression is additive. The studies based on this assumption use linear regression to relate regulatory sequence elements to gene expression values [12, 60]. These approaches, however, cannot identify synergistic regulatory element combinations that control gene expression patterns. Algorithms have been developed to model the synergy between two transcription factors that bind to sites located anywhere in the upstream region [87] or sites that are spatially close to each other [39]. Beer and Tavazoie [5] presented an approach which utilizes AND, OR and NOT logic to capture combinatorial gene regulation. This method is not only able to infer combinatorial rules that involve the presence of binding sites for more than two regulators, but it also includes constraints on motif strength, orientation, relative position and additional copies of the motifs. Although the methods that model

combinatorial effects of the motifs have appealing properties, their drawback is their inability to cope with uncertainty in the transcription factor binding sites that are identified. The robustness of the method in the face of uncertainty is important, as non-functional transcription factor binding sites can be readily found throughout the genome, including promoters [121]. We present an approach which is both able to capture the combinatorial nature of gene regulation and to incorporate uncertainty about the functionality of putative regulatory sequence elements. Our probabilistic method, which is based on noisy threshold models, extends the earlier methods that infer combinatorial rules in two directions. First, we consider a larger class of Boolean functions, Boolean threshold functions, to capture combinatorial effects. Second, the regulatory sequence elements contribute to the regulation of a gene through hidden variables that capture the probability that a sequence element is functional; thus, the method is able to cope with non-functional regulator binding sites.

The methods that use both expression and genome sequence data to explain gene regulation can be divided into three groups depending on the way in which the two sources of data are combined. The first group includes the methods that first cluster genes on the basis of their expression patterns and then search for putative motifs in the upstream regions of the genes in each cluster [10, 18, 39, 107]. The methods in the second group work in the opposite direction, first identifying a set of candidate motifs and then trying to explain RNA expression using these motifs [12, 52, 87]. Finally, the algorithms in the last group use both sources of data together. These methods use one or more iterations of the following procedure: first, genes are clustered or grouped according to their expression data, then the search for motifs in the upstream regions of the coexpressed genes is performed, and, then, the motifs identified are used to build models that predict the expression pattern of the gene (see e.g. [5, 99]). While methods in the third group are potentially more powerful than methods in the other two groups, they require more information about the gene function than is available for *Plasmodium*. Specifically, they require a fine-grained clustering based on expression data and annotation. For *Plasmodium*, only 40% of the genes are annotated [36] and only a few gene expression datasets are available. In contrast to yeast (the organism on which the approaches in group 3 are tested), gene expression in *P. falciparum* is difficult to induce [95]. The approach we present belongs to the second group of methods. We do not use RNA expression data while searching for putative regulatory motifs; therefore, the accuracy of the models in predicting the gene expression pattern is an unbiased measure of the soundness of the mod-

els learned. An unbiased measure to evaluate the soundness of the results allows us to validate the results even though knowledge about the gene regulation processes is not available. This property of our method is critical given how little is known about gene regulation in *P. falciparum*.

This chapter investigates a number of questions about gene regulation in *P. falciparum* in the intraerythrocytic cycle. In section 5.2, we describe our approach which uses the noisy threshold models to identify regulatory sequence elements involved in gene regulation, to determine the predictive power of additional constraints and copies of these regulatory sequence elements and to examine the intriguing gene regulation pattern we find. We also propose a comparative genome analysis approach to identify potential transcription factors that bind to the sequence elements that play a regulatory role in gene expression. Section 5.3 presents our findings, and the most important findings are interpreted in Section 5.4.

5.2 Methodology

Our approach to infer regulatory modules from genome sequence and RNA expression data is shown in Figure 5.1. The underlying assumption in the approach is that genes within a cluster share common regulatory mechanisms. We start with a preprocessing step, where we use a motif-finding algorithm to identify putative regulatory sequence elements and we cluster genes according to their expression profiles. Then, for each cluster of genes that exhibited significant changes in their expression, we learn noisy threshold models, which model combinatorial effects of gene regulators, and, given the putative regulatory sequence elements for a gene, classify the gene as belonging to the cluster or not.

The global structure of a noisy threshold model can be seen in Figure 5.1; it expresses the idea that causes (regulatory sequence elements) influence a given common effect (gene membership to a gene expression cluster) through hidden variables and a Boolean threshold function. All variables in this model are binary; the positive state of a cause variable corresponds to either the absence or presence of the motif, and the positive state of the effect variable represents that the gene belongs to the cluster. Hidden variables are considered to be a contribution of their parent variables, regulatory sequence elements, to the gene expression pattern; the absent causes do not contribute to the effect. The Boolean threshold function τ_k returns true when there are at least k trues among the hidden variables.

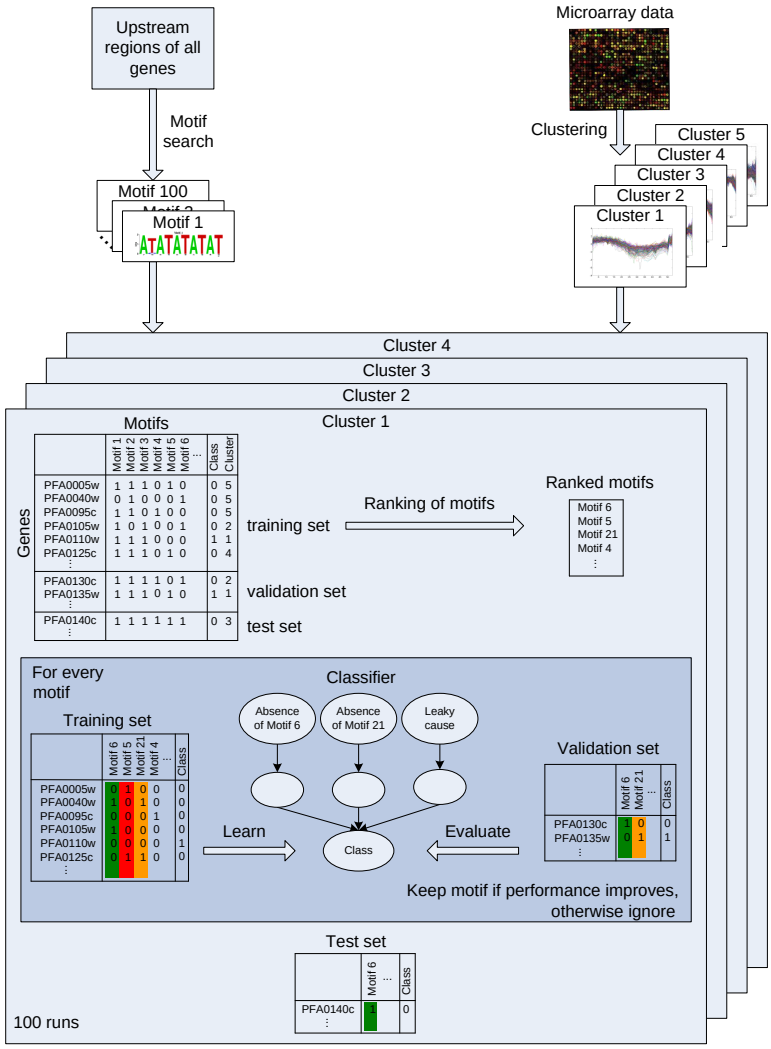


Figure 5.1: Overview of the proposed approach. After data preprocessing is completed, we perform 100 runs of two error estimation methods, cross-validation and bootstrap, for every cluster. Every run starts with motif ranking, where motifs are ranked based on mutual information scores computed between the motifs and the class. An iterative greedy procedure used to learn a noisy threshold model adds the next highest ranked motif (orange color) to the current model. A motif that improves classification performance (green color) is kept in the model, a motif that does not improve the classification performance (red color) is removed.

The commonly used OR and AND functions are the extremes of a spectrum of threshold functions: the OR function is a threshold function τ_k with $k = 1$ and the AND function is a threshold function τ_k where k equals the number of causes in the model.

5.2.1 Finding regulatory motifs

We extracted the DNA sequence 1000 bp upstream from the initiation codon of all *P. falciparum* protein coding genes using the latest version of PlasmoDB [2]. In instances where the upstream regulatory region overlapped with another open reading frame, we extracted only the sequence between the open reading frames. To find over-represented motifs, the extracted sequences were analyzed using the AlignACE program [51]. We set the GC background parameter to 0.13 (the fractional GC background for these regions), the number of columns to align to 10 and the number of expected sites to 5. Sequence logos of the motifs were generated using the WebLogo program [21].

5.2.2 Clustering of the RNA expression data

We used a *P. falciparum* 3D7 strain RNA expression data set [9]. We downloaded data that were normalized and median-centered and we only used data for those oligonucleotides that have a corresponding open reading frame assigned from PlasmoDB. We discarded the genes for which more than 20% of the measurements were missing. A number of open reading frames had more than one oligonucleotide measured; we averaged the measurements of these open reading frames. After the data had been $\log_2$ transformed, we imputed missing values using the weighted K-nearest neighbours method. We chose to use this data imputation method as it has been shown to provide a more robust and sensitive missing value estimation in microarray data than a singular value decomposition based method or the commonly used row average method [110]. The weighted K-nearest neighbours method uses a weighted average of values from the K genes closest to the gene of interest as an estimate for the missing value. Based on the results reported in [110], we chose the value of K to be 15 and the Euclidean distance as a metric for gene similarity.

We used the K-means algorithm with random initializations to cluster the genes according to their RNA expression data. Since the K-means al-

gorithm is known to sometimes get stuck in a local optimum, we ran the algorithm 10 times for each number of clusters. To select the optimal number of clusters we used the so-called C-index [50], which has been shown to outperform 13 other indices for determining the number of clusters in binary data sets when the data are clustered using the K-means algorithm [28].

5.2.3 Learning noisy threshold models

We split the data into training, validation and test sets. The training set was used to rank the motifs and learn a noisy threshold model, the validation set was used to choose the model parameters, and the test set was used to evaluate the classification performance of the model.

We used an iterative greedy approach to learn a noisy threshold model that separates the genes in cluster i from all other genes based on motif absence/presence. First, we ranked all motifs based on their mutual information scores, where the mutual information measures the mutual dependence of the variable M that represents a motif and the class variable C and is defined as

$$I(M;C) = \sum_{m \in M} \sum_{c \in C} \Pr(m,c) \log \frac{\Pr(m,c)}{\Pr(m)\Pr(c)}.$$

Variables M and C are binary, their true values denote the presence of at least one binding site for a motif in the upstream region of a gene and the belonging of the gene to the cluster for which a model is learned, respectively. Then, to learn a noisy threshold model, we started from a model containing no causes except a default so-called leaky cause [45] and iteratively added the next highest ranked motif. If the new model did not have a higher classification accuracy on the validation set than the previous model, the motif was removed from the model. For each newly added motif, we evaluated two models, a model with the interaction function τ_k and a model with the interaction function τ_{k+1}, where τ_k is the interaction function from the model with the highest classification accuracy in the previous iteration. To learn the probabilities of hidden variables in a noisy threshold model, we ran 10 iterations of the EM algorithm described in Chapter 3, computed the classification accuracy on the validation set after each iteration and chose the number of iterations that provided the highest classification accuracy.

To solve the problem of unbalanced data (different class size), we added as many copies of every gene from the smaller class as was needed for this class to amount for at least half of the genes in both classes.

5.2.4 Evaluation of the models learned

We used two error estimation methods, cross-validation and bootstrap, to evaluate the models learned. The cross-validation scheme was used to examine the predictive performance of the models, whereas the bootstrap approach was used to evaluate the reliability of the model parameters, the threshold function values and the motifs that were selected as model features. We performed 100 runs of both error estimation methods.

To compare two models learned under different assumptions as well as to compare our model to a classifier that assigns all genes to the bigger class, we used the altered form of the exact version of the McNemar's test [96]. Let n_1 and n_2 be the number of genes from the bigger and the smaller class, respectively, for which the classifiers produce different outputs, and let c be the number of copies of genes from the smaller class used to solve the problem of unbalanced data. We have two series of tests, one for the bigger class and the other for the smaller class, that can be respectively represented by two independent variables that follow the distributions $\mathrm{B}(n_1, \frac{1}{2})$ and $c\mathrm{B}(n_2, \frac{1}{2})$, where $\mathrm{B}(n, p)$ denotes the binomial distribution. The sum of the two variables can be approximated by a normal distribution with parameters $\mu = \frac{n_1+cn_2}{2}$ and $\sigma^2 = \frac{n_1+c^2n_2}{4}$. Let $\Phi_{\mu,\sigma^2}(x)$ denote the cumulative distribution function of the normal distribution. Then, under the null hypothesis that the two classifiers perform equally well, we compute the p-value

$$p = 1 - \Phi_{\mu,\sigma^2}(s),$$

where s is the number of cases where the output of the examined classifier was correct, while the output of the reference classifier was wrong.

Under the assumption that all motifs are equally likely to be chosen as features of the model, we used the following test to identify the motifs that are significant for the classification of the genes. From the results of the bootstrap approach, we estimate $\hat{\boldsymbol{p}} = (\hat{p}_1, \ldots, \hat{p}_{100})$, where

$$\hat{p}_i = \frac{\text{number of motifs in the model } i}{\text{total number of motifs}}$$

is the estimate of the probability that a motif will randomly be chosen as a feature in the model i. Letting Y denote the number of times a motif is chosen as a feature in model, our null hypothesis states that Y follows a Poisson binomial distribution (discrete probability distribution of the number of successes in a sequence of n independent experiments, each of which yields success with a different probability) with probabilities $\hat{\boldsymbol{p}}$ and number of trials $n = 100$. Let $\mathrm{P}(n, \hat{\boldsymbol{p}})$ denote the Poisson binomial distribution, then a motif is called significant if its p-value $p = \sum_{i=t}^{100} \mathrm{P}(i, \hat{\boldsymbol{p}})$ is less than the significance level of 0.05 corrected for multiple motifs tested, where t is the number of models the motif appears in.

5.2.5 Examining constraints and copies of the binding sites of the motifs

To examine constraints on the binding sites of the motifs, we introduced three binary variables that represent constraints on orientation, location with respect to ATG and functional depth. We computed mutual information scores between each of the variables and the class variable, where the true values of the variables were these: forward orientation, location within a given interval and fractional score higher than the given score. Functional depth is defined as the ratio between a binding site score and the maximum motif score. The scores are the sums obtained from the position weight matrix where the element $m_{i,b}$ of the matrix is computed as

$$m_{i,b} = \ln \frac{N_{i,b} + \Pr(b)}{(N+1)\Pr(b)},$$

where $N_{i,b}$ is the number of bases of type b aligned at position i, N is the number of aligned binding sites, and $\Pr(b)$ is the background frequency for base b.

The mutual information scores of the constraints on the binding sites of a motif were included in the framework of learning the noisy threshold model only if the motif has been tested and improved the classification accuracy of the model. Once the mutual information scores of the constraints for the newly added motif have been included, all mutual information scores were sorted in descending order and the highest ranking motif or constraint on the binding sites of a motif was tested at the next iteration. The procedure was continued until all motifs and constraints on the binding sites of the motifs selected as model features were tested.

The experiments to test whether information about additional binding sites of a motif improves the classification performance were performed in a very similar manner as those testing the usefulness of the information about the constraints on the binding sites of a motif. The mutual information score of the $(l+1)$-th binding site of a motif was included in the framework only if the l-th binding site of the motif has been tested and improved the classification accuracy of the noisy threshold model.

5.2.6 Identifying potential transcription factors binding to the motifs

To identify potential transcription factors binding to the motifs, we used comparative genome analysis, which is based on the fact that sequence similarity across species might reflect functional similarity. Identification, which was done separately for each significant motif, involved three steps. Firstly, we used STAMP [74], a web tool for exploring DNA-binding motif similarities, to find a number of the closest matches for a given motif. Secondly, for each match found, we checked whether the database where the motif is stored reports a transcription factor that binds to it. Finally, if the transcription factor is known, we used NCBI BLAST [1, 98] to find the most similar protein sequences in the *P. falciparum* protein database.

5.3 Results

5.3.1 Inferred significant motifs

Since the average length of upstream sequences that contribute to regulation is not known, we tested two sets of upstream sequences with different lengths: 1000 and 1500 bp. The models learned from 1000 bp upstream sequences showed slightly better classification performance; therefore, we report and analyze the results obtained using the set of 1000 bp upstream sequences.

We chose the number of clusters to be 5, as the C-index [50] curve had an 'elbow' at this value. The clusters are comparable to the four characteristic stages of intraerythrocytic parasite morphology discussed by Bozdech et al. [9], as the vast majority of genes induced in every one of the stages belongs to one of four clusters. Cluster 5 is a cluster of genes whose

Table 5.1: Summary of the clusters and classification accuracy of the noisy threshold models learned to explain them. The p-values are computed for the null hypothesis that the noisy threshold models are just as good as a baseline classifier which assigns all genes to the bigger class.

Cluster	Number of genes	Corresponding stage	Classification accuracy (%)	Baseline accuracy (%)	p-value
1	144	early ring (ER)	58.5	50.4	$3.4 \cdot 10^{-3}$
2	1033	ring/early trophozoite (RET)	60.8	52.5	$9.7 \cdot 10^{-15}$
3	985	trophozoite/early schizont (TES)	58.6	50.9	$4.1 \cdot 10^{-11}$
4	329	schizont (S)	60.5	50.8	$2.2 \cdot 10^{-6}$
5	1344	-	-	-	-

expression is more or less constant throughout the intraerythrocytic stage. The correspondence between the characteristic stages and the clusters is given in Table 5.1. From now on, we refer to the clusters by the names of the corresponding characteristic stages.

We used 100 motifs that AlignACE found to learn the models for the first four clusters, i.e. the clusters of genes whose expression changed throughout the intraerythrocytic stage. The classification accuracy of the noisy threshold models learned using the cross-validation procedure is reported in Table 5.1. Table 5.1 shows that we attain an accuracy of around 60% for each of the stages with little difference in accuracy between the stages. The predictability of the gene expression from upstream sequence elements is around the same as the predictability for *Saccharomyces cerevisiae* reported in [125]. All 39 motifs that were selected as features of the classifier in a significant number of bootstrap runs are shown in Figure 5.2.

The percentage of significant motifs that are present in upstream regions gradually increases in genes expressed in different stages throughout the intraerythrocytic cycle. Not surprisingly, the increase is strongest in the motifs that were significant for predicting the expression of genes in three or four clusters. It is interesting to note that the percentage of the motifs that predict gene expression positively correlates with the erythrocytic malaria parasites' total parasite protein content [64] (shown in Figure 5.3) and mRNA half-lives [102].

Figure 5.4 summarizes the parameters of the noisy threshold models learned

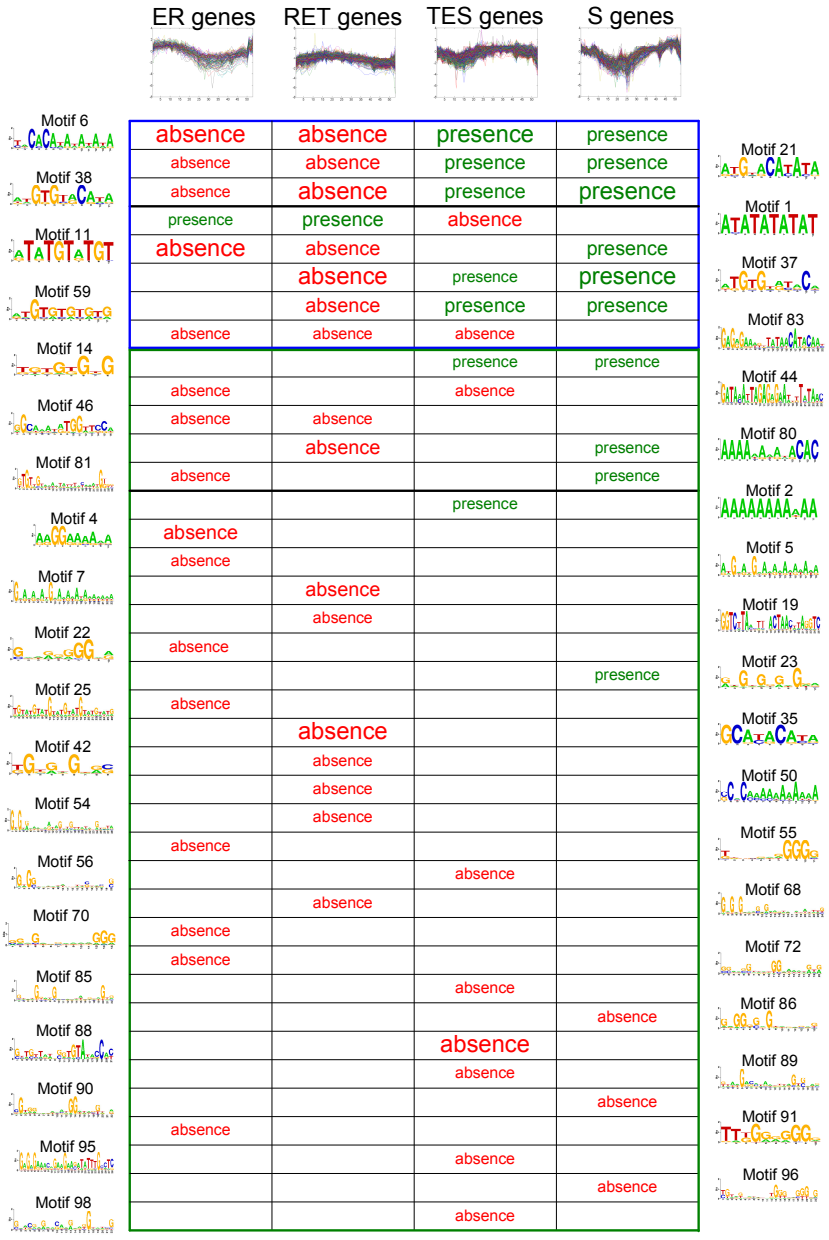


Figure 5.2: Motifs significant for predicting the expression of genes in different clusters. The motifs are ordered from top to bottom in terms of how often they appear as a feature explaining membership to an expression cluster. Font size indicates relative importance of a motif: large, medium and small font sizes indicate a motif selected as a feature in at least 75 bootstrap runs, less than in 75 but at least in 50 bootstrap runs and less than in 50 but in a significant number of bootstrap runs, respectively.

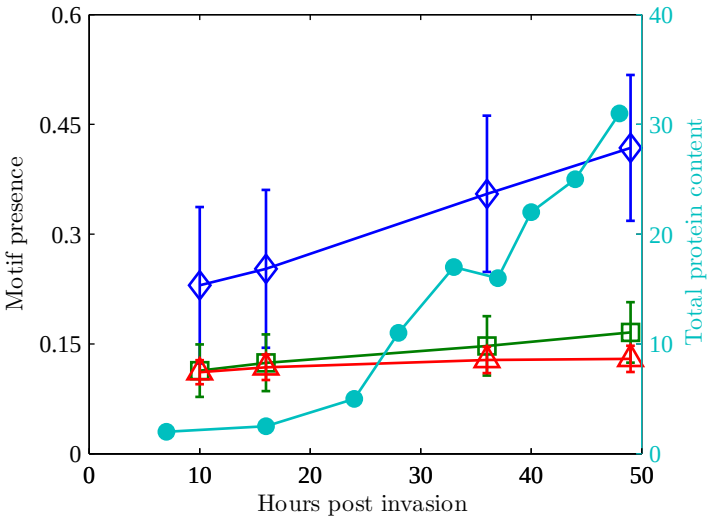


Figure 5.3: Both the total protein content (cyan) and the percentage of the motifs that predict gene expression (motifs selected in: 0 classifiers (red), 1 or 2 classifiers (green), and 3 or 4 classifiers (blue)), gradually increase with parasite maturation. The mean of the percentage of the motifs present in the genes in a cluster is plotted at the hour when most genes in the cluster show peak expression. The error bars represent the standard deviation of the mean.

using the bootstrap procedure. The median of the threshold function values for the clusters of TES and S genes is approximately four, meaning that presence of four or more functional regulatory elements predicts expression of genes in these clusters. These results match the findings of Van Noort and Huynen [114], who reported that most *Plasmodium* genes have between three to seven different regulatory elements in their upstream regions. The median of the standard deviation of the probabilities of hidden variables provides evidence that noisy threshold models capture the probability that a sequence element is functional. There is little variation in the probabilities of the hidden variables in the models where cluster membership is explained by motif absence, which allows us to conclude that motifs have very similar explanatory power. However, there is much more variation in the probabilities of the hidden variables in the models where cluster membership is mostly explained by motif presence; this can be explained by different rates of functional binding sites for different motifs.

To verify that the models represent gene regulatory mechanisms and could not be learned from sequence elements found just anywhere in the genome,

we learned models to predict gene expression using putative regulatory sequence elements found in 1000 bp downstream sequences. The models for all four clusters were not significantly better than a baseline classifier which assigns all genes to the bigger class (results not shown).

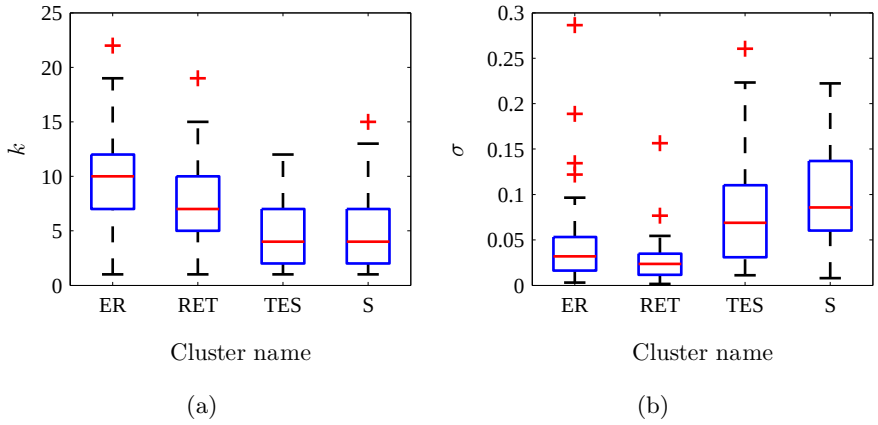


Figure 5.4: Summary of the parameters of noisy threshold models learned in 100 bootstrap runs: box and whisker plots of the threshold functions τ_k (a), and of the standard deviation σ of the probabilities of hidden variables in a model (b). Median (horizontal line), quartiles (box), 5% and 95% confidence intervals (whiskers) and outliers are indicated.

5.3.2 Pattern of present/absent motifs

Figure 5.2 reveals a distinct pattern in that all motifs for the earlier ER and RET stages, with the exception of Motif 1, explain membership by their absence, while many of the motifs for the later TES and S stages explain membership by their presence. Interestingly, the motifs that break this pattern, with the exception of Motif 1, are found in a small number of genes (from 1 to 5% of the genes), in sharp contrast to the other significant motifs, which are more common. We put this observation to the test by learning noisy threshold models in which only either the presence or the absence of the motifs could be selected as causes. As Table 5.2 shows, the models that follow the pattern were about as good as the original models, whereas models that break the pattern did not perform better than the baseline classifier.

Table 5.2: Classification accuracy of models in which only either the presence or the absence of the motifs were selected as causes. The p-values are computed for the null hypothesis that the noisy threshold models are just as good as a baseline classifier which assigns all genes to the bigger class.

Cluster	Accuracy of models following the pattern (%)	p-value	Accuracy of models breaking the pattern (%)	p-value
ER	59.5	$1.1 \cdot 10^{-3}$	50.9	0.44
RET	61.3	$1.1 \cdot 10^{-15}$	52.4	0.56
TES	58.5	$5.3 \cdot 10^{-11}$	51.1	0.39
S	59.9	$6.2 \cdot 10^{-6}$	49.3	0.85

5.3.3 Additional information on the regulatory sequence elements

To test whether additional information about the regulatory sequence elements is useful for predicting gene expression, we performed two sets of experiments. In the first set of experiments, we learned models where constraints on motif's orientation, location with respect to ATG and functional depth were added if they improved classification performance on the validation set. Likewise, in the second set of experiments, we learned models where information about additional copies of a motif was included into the model if it improved the classification performance on a validation set.

The classification accuracy of the models with constraints on motifs and p-values for the null hypothesis that the models perform equally well as the original models are shown in Table 5.3. Even though there is a slight tendency for a few motifs in the RET cluster to have a positional preference - e.g. in a number of models, Motif 6 is constrained to 250 - 500 bp upstream, and Motif 1 is constrained to 500-1000 bp - this seems to be circumstantial. These constraints are not preserved in the models for the other clusters and the p-values corrected for multiple testing (corrected for eight models learned in this section) are not significant. Yuan et al. [125] reach the same conclusion in their paper, where the authors showed that the orientation and position information for predicted transcription factor binding sites in *Saccharomyces cerevisiae* do not help in predicting coexpression of genes.

The right side of Table 5.3 lists the classification accuracy of the models

with additional binding sites and p-values for the null hypothesis that the models perform equally well as the original models without information about the number of motif binding sites per gene. The results suggest that, even if some of the genes have multiple functional copies of motifs in their upstream regions, additional copies of motifs do not help to predict gene expression.

Table 5.3: Classification accuracy of models with additional constraints and models with additional binding sites.

Cluster	Accuracy of original models (%)	Accuracy of models with constraints (%)	p-value	Accuracy of models with additional binding sites (%)	p-value
ER	58.5	56.4	0.90	59.1	0.25
RET	60.8	61.4	0.22	61.6	0.12
TES	58.6	60.2	0.01	58.9	0.31
S	60.5	61.4	0.21	60.3	0.60

5.3.4 Correspondence to functionally tested sequence motifs

In Figure 5.2, we presented a list of 39 motifs that are deemed significant by our approach. We obtained support for the hypothesis that many of these motifs are functional by using STAMP [74] to compare functionally tested *Plasmodium* sequence motifs [13, 17, 24, 40, 48, 66, 77, 92] with the putative regulatory sequence motifs we used for learning the models. The results of this comparison, presented in Figure 5.5, are encouraging as the vast majority of the best matches were the putative regulatory sequence motifs found to be significant for predicting gene expression patterns. A few interesting observations that emerge from this comparison are discussed further.

The TATA box sequence (TATAA) which was shown to be bound by the TATA box binding protein [92] is a part of two motifs, i.e. Motif 83 and Motif 44. It is interesting to note that in both motifs the TATAA sequence is preceded by the GAGAGAA sequence, which has been reported as a putative enhancer element in *P. falciparum* [40], and in both cases the distance between these two elements is three nucleotides.

Even though the TCATA sequence was not found in any of the motifs, the three closest matches, Motif 11, Motif 21 and Motif 35, had the same

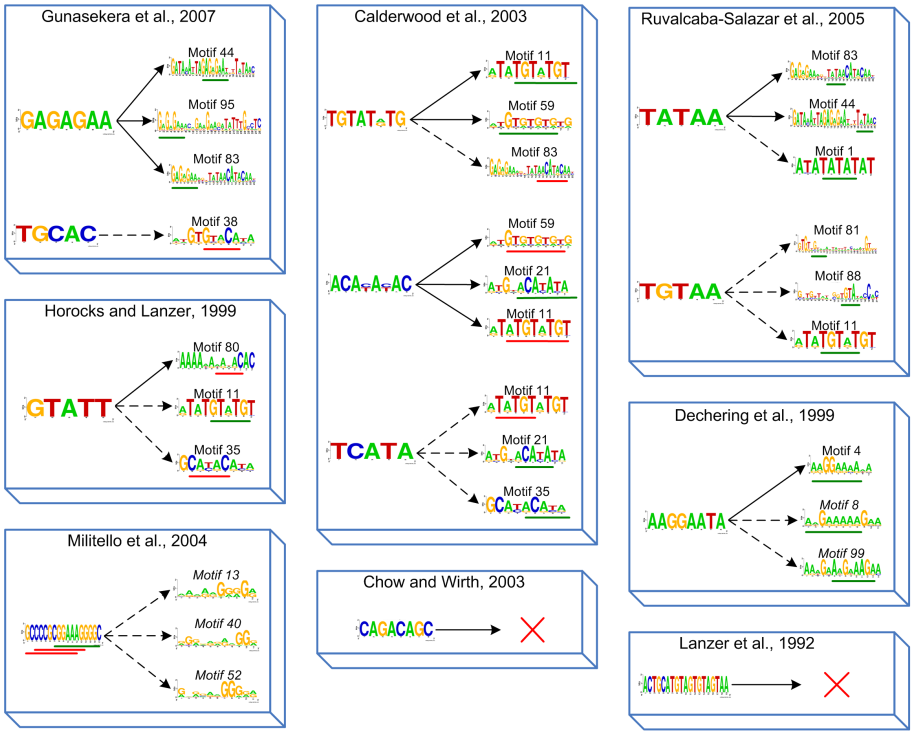


Figure 5.5: We report up to three best matches for a functionally tested sequence motif if the E-value of the match is less than 10^{-2}. Dashed arrows indicate weak matches whose E-values are more than 10^{-3} for functionally tested motifs that are up to 5 bp long and more than 10^{-5} for functionally tested motifs that are longer than 5 bp; solid arrows indicate the better matches. The locations of the alignments are underlined by green (forward alignment) and red (reverse alignment) lines in longer sequences of the matches. The names of the significant motifs are written in roman, the names of the motifs that we did not find to be significant are written in italic.

TCATA sequence-like sequence, the ACATA sequence.

We did not find close matches to the dual palindromic G-boxes GCCCCGCGGAAAGGGGC, which were shown to activate gene expression in *Plasmodium* species [77]. Differently from the matches to the other functionally tested sequences, all three closest matches to the dual palindromic G-boxes, Motif 13, Motif 40 and Motif 52, are motifs that were found to be not significant for predicting gene expression pattern.

We do not report the results of the comparison with the functionally tested sequence elements reported in [84, 89] as these long sequence elements are extremely AT-rich, what makes it difficult to evaluate the matches obtained.

5.3.5 Potential transcription factors that bind to the motifs

Comparative genome analysis (see Figure 5.6), in which we searched nine databases of both eukaryotic and prokaryotic transcription factor binding motifs [7, 41, 46, 59, 73, 75, 97, 106, 123], resulted in the disclosure of ten potential transcription factors in *P. falciparum*. Our criterion for labelling a gene as a potential transcription factor was the gene being found as a match via more than one transcription factor binding motif from other organisms. The rationale for this criterion was based on the fact that our approach ended up with less than sixty genes in *P. falciparum* whose E-value (a parameter that represents the number of times this match or a better one would be expected to occur purely by chance in a search of the entire database) was less than or equal to 1; therefore, the probability of a gene being found by chance via different matches is very small. Nine genes were found repeatedly using the comparative genome analysis approach: *PFL0465c*, *PF14_0175*, *PF13_0198*, *PF13_0072*, *PF11_0294*, *MAL3P7.34*, *PFB0540w*, *PF10_0143* and *MAL13P1.176*. The analysis showed one more potential transcription factor, *PF14_0316*, putative DNA topoisomerase II, which did not meet the criterion above, but was an identical match to the gene in the fruit fly; furthermore it was found via the closest matches for two very important motifs in predicting gene expression patterns, Motif 6 and Motif 11. Figure 5.6 shows the alignments which produced the potential transcription factors; for E-values of the motif and protein sequence matches, see Supplementary Tables 5.4-5.13.

The natural question to ask is how many of these genes appeared as significant matches only because they are paralogs of a true transcription

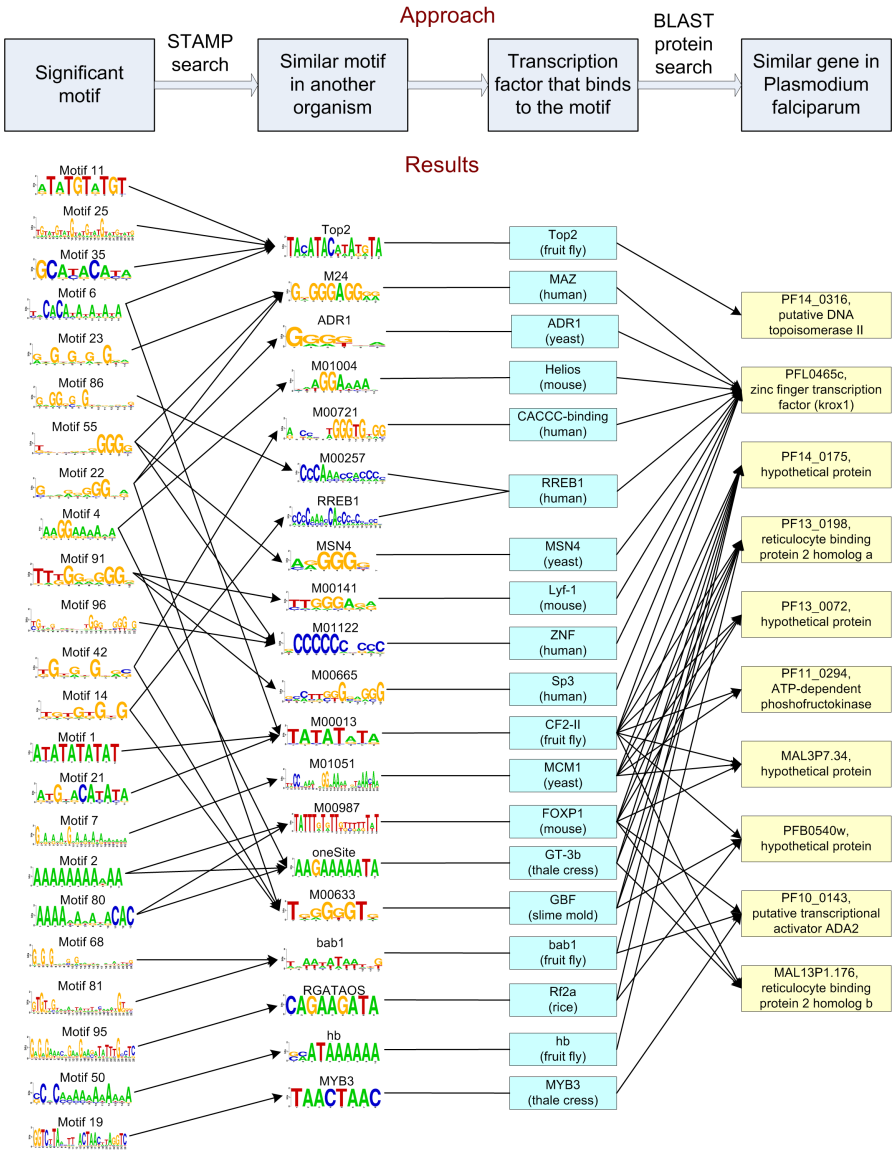


Figure 5.6: Comparative genome analysis to identify potential transcription factors: approach and results. Most of the ten genes identified as potential transcription factors show no significant change in their expression during the intraerythrocytic cycle; eight of them in fact belong to the silent cluster 5. The expression of the other two potential transcription factors, *PF13_0198*, reticulocyte binding protein 2 homolog a, and of *MAL13P1.176*, reticulocyte binding protein 2 homolog b, peaks during the schizont stage.

factor. To answer this question, we used NCBI BLAST [1, 98] to examine how similar the protein sequences of these potential transcription factors are. Three of the genes, *PF14_0316*, *PFL0465c* and *PF11_0294*, are not similar to any other gene from the list. Two of the genes, *PF13_0198* and *MAL13P1.176*, are paralogs of each other, and both of them are similar to the hypothetical protein *PF14_0175* with E-values of $5 \cdot 10^{-6}$ and $2 \cdot 10^{-8}$, respectively. Even though the protein sequences of the other genes have some similarities between themselves and with the two paralogs, none of the sequences are closely related as there is only one match whose E-value is less than 10^{-2}: *MAL3P7.34* is similar to *PFB0540w* with an E-value of $6 \cdot 10^{-5}$.

5.4 Discussion

Previous bioinformatics approaches have yielded some information on transcription regulatory elements and transcription factors in *P. falciparum*. Callebaut et al. [14] predicted general transcription factors associated with RNA polymerase II. Several approaches searched for regulatory elements in the upstream sequences of a gene family [77] or clusters of co-expressed genes [114, 124].

Differently from other bioinformatics approaches applied to *P. falciparum*, our method is able to model the logic behind gene regulation and to incorporate uncertainty about the functionality of putative regulatory sequence elements. The classification accuracy of the noisy threshold models, which is an unbiased measure of the soundness of the models, allowed us to test a number of different models and, consequently, different hypotheses about gene regulation. Our main findings are as follows. First, we report a prioritized list of thirty nine regulatory motifs relevant for gene regulation and we show that prevalence of these motifs increases with progression through the developmental cycle. Second, we show that several factors (other than DNA sequence of the motif) are unlikely to contribute to gene regulation. Third, we provide a list of ten potential transcription factors with their associated binding motifs.

Of the thirty nine significant motifs eleven are implicated in the regulation of genes expressed in the second phase of the asexual development in the blood cell (motifs that have a "presence" label in column 3 or 4 in Figure 5.2). The second phase consists of the final 18 hours of the intraerythrocytic cycle, in which nuclear division is followed by merozoite

formation and release [33]. Given that hundreds of genes are involved in nuclear division, it is likely that these motifs contribute to the regulation of the expression of the genes involved in mitosis. Supporting evidence for this hypothesis could be our identification of *PF14_0316* as a transcription factor. *PF14_0316* is annotated as a putative topoisomerase-II in PlasmoDB 5.4 and it has been shown by Kelly et al. [61] that treatment of asexual stage *P. falciparum* parasites with the topoisomerase-II inhibitor etoposide results in chromosomal cleavage.

There are two peculiar findings of our approach. First, cluster membership of genes expressed in the first two stages is predicted solely by the absence of motifs (with exception of Motif 1, see Figure 5.2). Second, we found that the relative abundance of motifs upstream of cluster 5 genes, the cluster of genes that do not significantly change expression during the intraerythrocytic cycle, was halfway between the abundance of motifs upstream of the early and late stages genes. One possible explanation is that we found no evidence of regulation of the genes expressed in the initial phase of the intraerythrocytic cycle, in which principal modifications of the host cell occur that allow the parasite to transport molecules in and out of the cell, to prepare the surface of the red blood cell to mediate cytoadherence and to digest the cytoplasmic contents in its food vacuole. For this hypothesis, the number of motifs not being the same or less than in genes of the first stage could be explained by these motifs being relevant in different combinations in other life cycle stages. A second possible explanation for these findings could be that the lack of regulatory motifs is a kind of gene regulation. Motif 4 seems to be strong evidence for this explanation as it is found in 42% of the genes of the ER cluster as compared to 63% of the genes in the other four clusters; furthermore, the percentage of Motif 4 almost does not vary throughout these four clusters. For this hypothesis, the number of motifs in cluster 5 is a baseline number of motifs found in the genes. We did not find any evidence for a third possible explanation - that a fraction of the genes in cluster 5 should belong to one of the other 4 clusters.

Our model allowed us to test two sets of upstream sequences with different lengths. We found that models learned from the set of 1000 bp upstream sequences showed better classification performance than models learned from the set of 1500 bp upstream sequences. This result might look surprising, given that a cis-acting sequence element has been found as far away as 1600 bp upstream [84]. However, genes having regulatory elements that are far from the translation start site are probably an exception since 43%

of the genes have another open reading frame starting less than 1500 bp upstream, and, additionally, 21% of the genes share a region where the distance between the translation start sites of the two genes is less than 3000 bps.

Our second main finding is the irrelevance of ancillary information for predicting gene expression. Our model allowed us to test whether extra information besides the DNA sequence of the motif, namely, score, orientation, location and multiplicity of the motif, makes a significant contribution to explaining cluster membership. In our experiments, this additional information about the motifs did not improve the prediction of gene expression. This does not imply, however, that these motif properties are biologically irrelevant.

Our third main finding are ten potential transcription factors and their associated DNA binding motifs. Although it is too optimistic to expect that all ten are involved in transcriptional regulation, we believe that some of them can be true transcription factors, given that one of them, *PFL0465c*, is a zinc finger transcription factor and another, *PF10_0143*, is a putative transcriptional activator. There are three reasons why we think it is very unlikely that genes *PF13_0198* and *MAL13P1.176* are transcription factors: (1) they are transmembrane proteins localising to the rhoptry organelles; (2) both of them are similar to the hypothetical protein *PF14_0175*, which is similar to a number of transcription factors in other organisms; (3) their expression pattern is different from that of the other potential transcription factors. The fact that eight out of the ten potential transcription factors show no significant change in their expression during the intraerythrocytic cycle suggests that transcription factors specific to the intraerythrocytic cycle are likely to be regulated post-transcriptionally.

The analysis to identify potential transcription factors also revealed that many of the significant motifs were similar to the same motifs in other organisms, and, thus, similar among themselves. This allows us to hypothesize that at least some of the motifs are bound by transcription factors that bind a family of similar but distinct motifs.

Acknowledgments

We would like to thank Michael A. Beer, Saeed Tavazoie, Zbynek Bozdech and Mahony Shaun for their help and clarifications.

5.5 Supplementary material

Table 5.4: Alignments that identified **PF14_0316**, putative DNA topoisomerase II, as a potential transcription factor in *P. falciparum*.

Statistically significant motif	Similar motif in another organism	E-value of the motif match	Transcription factor that binds to this motif	E-value of the protein match
Motif 11	Top2	$3 \cdot 10^{-14}$		
Motif 25	Top2	$3 \cdot 10^{-11}$		
Motif 6	Top2	10^{-9}		
Motif 35	Top2	10^{-7}	Top2 (fruit fly)	0.0

Table 5.5: Alignments that identified **PFL0465c**, zinc finger transcription factor (krox1), as a potential transcription factor in *P. falciparum*.

Statistically significant motif	Similar motif in another organism	E-value of the motif match	Transcription factor that binds to this motif	E-value of the protein match
Motif 23	M24	$3 \cdot 10^{-8}$		
Motif 22	M24	$2 \cdot 10^{-5}$		
Motif 55	M24	$5 \cdot 10^{-5}$	MAZ (human)	10^{-11}
Motif 1	M00013	10^{-12}		
Motif 21	M00013	$6 \cdot 10^{-8}$		
Motif 6	M00013	$9 \cdot 10^{-8}$	CF2-II (fruit fly)	$6 \cdot 10^{-8}$
Motif 4	M01004	$8 \cdot 10^{-8}$	Helios (mouse)	$2 \cdot 10^{-7}$
Motif 91	M00141	$3 \cdot 10^{-6}$	Lyf-1 (mouse)	$6 \cdot 10^{-6}$
Motif 42	M00721	$2 \cdot 10^{-5}$	CACCC-binding (human)	$6 \cdot 10^{-6}$
Motif 55	MSN4	$3 \cdot 10^{-4}$	MSN4 (yeast)	$2 \cdot 10^{-5}$
Motif 86	M00257	$4 \cdot 10^{-5}$		
Motif 14	RREB1	$7 \cdot 10^{-5}$	RREB1 (human)	$5 \cdot 10^{-4}$
Motif 22	ADR1	$5 \cdot 10^{-5}$	ADR1 (yeast)	$2 \cdot 10^{-2}$
Motif 91	M01122	10^{-6}		
Motif 96	M01122	$4 \cdot 10^{-5}$		
Motif 55	M01122	$3 \cdot 10^{-7}$	ZNF (human)	$2 \cdot 10^{-4}$
Motif 91	M00665	10^{-7}	Sp3 (human)	10^{-3}

Table 5.6: Alignments that identified hypothetical protein **PF14_0175** as a potential transcription factor in *P. falciparum*.

Statistically significant motif	Similar motif in another organism	E-value of the motif match	Transcription factor that binds to this motif	E-value of the protein match
Motif 80	M00987	$4 \cdot 10^{-9}$		
Motif 2	M00987	$2 \cdot 10^{-8}$	FOXP1 (mouse)	10^{-8}
Motif 7	M01051	$3 \cdot 10^{-9}$	MCM1 (yeast)	$2 \cdot 10^{-5}$
Motif 1	M00013	10^{-12}		
Motif 21	M00013	$6 \cdot 10^{-8}$		
Motif 6	M00013	$9 \cdot 10^{-8}$	CF2-II (fruit fly)	10^{-4}
Motif 14	M00633	$2 \cdot 10^{-7}$		
Motif 22	M00633	$7 \cdot 10^{-6}$		
Motif 42	M00633	10^{-4}	GBF (slime mold)	$3 \cdot 10^{-3}$
Motif 80	oneSite	10^{-8}		
Motif 4	oneSite	$3 \cdot 10^{-8}$		
Motif 2	oneSite	$8 \cdot 10^{-8}$	GT-3b (thale cress)	$8 \cdot 10^{-3}$
Motif 81	bab1	$2 \cdot 10^{-9}$		
Motif 68	bab1	$3 \cdot 10^{-7}$	bab1 (fruit fly)	$6 \cdot 10^{-2}$
Motif 95	RGATAOS	$8 \cdot 10^{-7}$	Rf2a (rice)	$9 \cdot 10^{-2}$

Table 5.7: Alignments that identified reticulocyte binding protein 2 homolog a **PF13_0198** as a potential transcription factor in *P. falciparum*.

Statistically significant motif	Similar motif in another organism	E-value of the motif match	Transcription factor that binds to this motif	E-value of the protein match
Motif 14	M00633	$2 \cdot 10^{-7}$		
Motif 22	M00633	$7 \cdot 10^{-6}$		
Motif 42	M00633	10^{-4}	GBF (slime mold)	$4 \cdot 10^{-8}$
Motif 80	M00987	$4 \cdot 10^{-9}$		
Motif 2	M00987	$2 \cdot 10^{-8}$	FOXP1 (mouse)	$2 \cdot 10^{-6}$
Motif 1	M00013	10^{-12}		
Motif 21	M00013	$6 \cdot 10^{-8}$		
Motif 6	M00013	$9 \cdot 10^{-8}$	CF2-II (fruit fly)	$4 \cdot 10^{-4}$
Motif 80	oneSite	10^{-8}		
Motif 4	oneSite	$3 \cdot 10^{-8}$		
Motif 2	oneSite	$8 \cdot 10^{-8}$	GT-3b (thale cress)	10^{-3}
Motif 50	hb	10^{-7}	hb (fruit fly)	$5 \cdot 10^{-2}$
Motif 7	M01051	$3 \cdot 10^{-9}$	MCM1 (yeast)	$9 \cdot 10^{-2}$

Table 5.8: Alignments that identified hypothetical protein **PF13_0072** as a potential transcription factor in *P. falciparum*.

Statistically significant motif	Similar motif in another organism	E-value of the motif match	Transcription factor that binds to this motif	E-value of the protein match
Motif 1	M00013	10^{-12}		
Motif 21	M00013	$6 \cdot 10^{-8}$		
Motif 6	M00013	$9 \cdot 10^{-8}$	CF2-II (fruit fly)	$6 \cdot 10^{-5}$
Motif 80	M00987	$4 \cdot 10^{-9}$		
Motif 2	M00987	$2 \cdot 10^{-8}$	FOXP1 (mouse)	$2 \cdot 10^{-3}$
Motif 7	M01051	$3 \cdot 10^{-9}$	MCM1 (yeast)	$3 \cdot 10^{-2}$

Table 5.9: Alignments that identified ATP-dependent phosphofructokinase **PF11_0294** as a potential transcription factor in *P. falciparum*.

Statistically significant motif	Similar motif in another organism	E-value of the motif match	Transcription factor that binds to this motif	E-value of the protein match
Motif 1	M00013	10^{-12}		
Motif 21	M00013	$6 \cdot 10^{-8}$		
Motif 6	M00013	$9 \cdot 10^{-8}$	CF2-II (fruit fly)	10^{-3}
Motif 7	M01051	$3 \cdot 10^{-9}$	MCM1 (yeast)	$5 \cdot 10^{-3}$

Table 5.10: Alignments that identified hypothetical protein **MAL3P7.34** as a potential transcription factor in *P. falciparum*.

Statistically significant motif	Similar motif in another organism	E-value of the motif match	Transcription factor that binds to this motif	E-value of the protein match
Motif 80	M00987	$4 \cdot 10^{-9}$		
Motif 2	M00987	$2 \cdot 10^{-8}$	FOXP1 (mouse)	$2 \cdot 10^{-2}$
Motif 7	M01051	$3 \cdot 10^{-9}$	MCM1 (yeast)	$3 \cdot 10^{-2}$
Motif 1	M00013	10^{-12}		
Motif 21	M00013	$6 \cdot 10^{-8}$		
Motif 6	M00013	$9 \cdot 10^{-8}$	CF2-II (fruit fly)	0.3

Table 5.11: Alignments that identified hypothetical protein **PFB0540w** as a potential transcription factor in *P. falciparum*.

Statistically significant motif	Similar motif in another organism	E-value of the motif match	Transcription factor that binds to this motif	E-value of the protein match
Motif 1	M00013	10^{-12}		
Motif 21	M00013	$6 \cdot 10^{-8}$		
Motif 6	M00013	$9 \cdot 10^{-8}$	CF2-II (fruit fly)	$4 \cdot 10^{-2}$
Motif 95	RGATAOS	$8 \cdot 10^{-7}$	Rf2a (rice)	$4 \cdot 10^{-2}$
Motif 14	M00633	$2 \cdot 10^{-7}$		
Motif 22	M00633	$7 \cdot 10^{-6}$		
Motif 42	M00633	10^{-4}	GBF (slime mold)	1

Table 5.12: Alignments that identified a putative transcriptional activator ADA2, **PF10_0143**, as a potential transcription factor in *P. falciparum*.

Statistically significant motif	Similar motif in another organism	E-value of the motif match	Transcription factor that binds to this motif	E-value of the protein match
Motif 81	bab1	$2 \cdot 10^{-9}$		
Motif 68	bab1	$3 \cdot 10^{-7}$	bab1 (fruit fly)	$9 \cdot 10^{-2}$
Motif 80	M00987	$4 \cdot 10^{-9}$		
Motif 2	M00987	$2 \cdot 10^{-8}$	FOXP1 (mouse)	0.2
Motif 19	MYB3	$2 \cdot 10^{-8}$	MYB3 (thale cress)	1

Table 5.13: Alignments that identified reticulocyte binding protein 2 homolog b **MAL13P1.176** as a potential transcription factor in *P. falciparum*.

Statistically significant motif	Similar motif in another organism	E-value of the motif match	Transcription factor that binds to this motif	E-value of the protein match
Motif 80	M00987	$4 \cdot 10^{-9}$		
Motif 2	M00987	$2 \cdot 10^{-8}$	FOXP1 (mouse)	0.1
Motif 1	M00013	10^{-12}		
Motif 21	M00013	$6 \cdot 10^{-8}$		
Motif 6	M00013	$9 \cdot 10^{-8}$	CF2-II (fruit fly)	0.3
Motif 80	oneSite	10^{-8}		
Motif 4	oneSite	$3 \cdot 10^{-8}$		
Motif 2	oneSite	$8 \cdot 10^{-8}$	GT-3b (thale cress)	0.7

Chapter 6

Conclusions and Further Research

In this thesis, we have studied symmetric causal independence models, a way to constrain the conditional probability tables for binary variables using symmetric Boolean functions.

The established connection between the conditional probabilities of the effect variable in the model and the Poisson binomial distribution enabled efficient probabilistic inference and parameter learning in symmetric causal independence models. We studied both exact and approximate inference methods using the computational scheme of the Poisson binomial distribution and compared their efficiency with that of the standard exact inference techniques. We used the computational scheme of the Poisson binomial distribution to develop a computationally efficient EM algorithm to learn the parameters in symmetric causal independence models. The investigation of the maxima of the log-likelihood function for symmetric causal independence models revealed that the log-likelihood for the noisy OR and the noisy AND models has only global maxima.

The competitive performance of the symmetric causal independence models present them as a potentially useful additional tool to the set of classifiers. We showed that the semantics of the models make them especially suitable for medical and genomic domains. We believe that noisy threshold models can be successfully applied to other eukaryotes for which it is

expected that genes are regulated by more than one transcription factor. Modelling transcriptional gene regulation in *Saccharomyces cerevisiae*, a widely used and one of the most studied model organism in science, is of particular interest.

This thesis has examined the problem of parameter learning in symmetric causal independence models, but the problem of learning an optimal interaction function has not been yet addressed. Efficient search in symmetric Boolean function space is a possible direction for future research.

Symmetric causal independence models studied in this thesis consist of binary variables. However, causal independence models do not have to be limited to binary variables. Researchers have proposed several schemes to generalize the noisy OR model to multivalued variables [27, 45, 105]. Extension of the framework of symmetric causal independence models to handle multivalued variables is another research challenge that remains to be addressed.

Publications by the Author

R. Jurgelenaite and T. Heskes. Learning symmetric causal independence models. *Machine Learning*, 71:133–153, 2008.

M.A.J. van Gerven, R. Jurgelenaite, B.G. Taal, T. Heskes, and P.J.F. Lucas. Predicting carcinoid heart disease with the noisy-threshold classifier. *Artificial Intelligence in Medicine*, 40:45–55, 2007.

R. Jurgelenaite, T. Heskes, and T. Dijkstra. Using symmetric causal independence models to predict gene expression from sequence data. In *Proceedings of the ECML-PKDD Workshop 'Data Mining in Functional Genomics and Proteomics: Current Trends and Future Directions'*, pages 67–78, 2007.

R. Jurgelenaite and T. Heskes. EM algorithm for symmetric causal independence models. In *Proceedings of the Nineteenth European Conference on Machine Learning*, pages 234–245, 2006.

R. Jurgelenaite and T. Heskes. Symmetric causal independence models for classification. In *Proceedings of the Third European Workshop on Probabilistic Graphical Models*, pages 163–170, 2006.

R. Jurgelenaite, P.J.F. Lucas, and T. Heskes. Exploring the noisy threshold function in designing Bayesian networks. In *Proceedings of AI-2005, the Twenty-fifth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 133–146, 2005.

R. Jurgelenaite and P.J.F. Lucas. Exploiting causal independence in large Bayesian networks. *Knowledge-Based Systems Journal*, 18:153–162, 2005.

R. Jurgelenaite and P.J.F. Lucas. Parameter estimation in large causal models. In *Proceedings of the Sixteenth European Conference on Artificial Intelligence*, pages 1037–1038, 2004.

Bibliography

[1] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402, 1997.

[2] A. Bahl, B.P. Brunk, J. Crabtree, M.J. Fraunholz, B. Gajria, G.R. Grant, H. Ginsburg, D. Gupta, J.C. Kissinger, P. Labo, L. Li, M.D. Mailman, A.J. Milgram, D.S. Pearson, D.S. Roos, J. Schug, C.J. Stoeckert Jr., and P.L. Whetzel. PlasmoDB: the Plasmodium genome resource. A database integrating experimental and computational data. *Nucleic Acids Research*, 31:212–215, 2003.

[3] D. Bamber. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12:387–415, 1975.

[4] R.C. Bast, D.W. Kufe, R.E. Pollock, R.R. Weichselbaum, J.F. Holland, and E. Frei, editors. *Cancer Medicine-5 Review*. B C Decker Inc., Ontario, 2000.

[5] M. A. Beer and S. Tavazoie. Predicting gene expression from sequence. *Cell*, 117:185–198, 2004.

[6] G. Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57: 33–45, 1962.

[7] C. M. Bergman, J. W. Carlson, and S. E. Celniker. Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster. *Bioinformatics*, 21:1747–1749, 2005.

[8] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[9] Z. Bozdech, M. Llinas, B. Pulliam, E.D. Wong, J. Zhu, and J. DeRisi. The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum. *PLoS Biology*, 1:85–100, 2003.

[10] A. Brazma, I. Jonassen, J. Vilo, and E. Ukkonen. Predicting gene regulatory elements in silico on a genomic scale. *Genome Research*, 8:1202–1215, 1998.

[11] J. Bremen. The ears of the hippopotamus: manifestations, determinants, and estimates of the malaria burden. *American Journal of Tropical Medicine and Hygiene*, 64:1–11, 2001.

[12] H.J. Bussemaker, H. Li, and E.D. Siggia. Regulatory element detection using correlation with expression. *Nature Genetics*, 27:167–171, 2001.

[13] M.S. Calderwood, L. Gannoun-Zaki, T.E. Wellems, and K.W. Deitsch. Plasmodium falciparum var genes are regulated by two regions with separate promoters,on upstream of the coding region and a second within the intron. *The Journal of Biological Chemistry*, 278:34125–34132, 2003.

[14] I. Callebaut, K. Prat, E. Meurice, J.P. Mornon, and S. Tomavo. Prediction of the general transcription factors associated with RNA polymerase II in Plasmodium falciparum: conserved features and differences relative to other eukaryotes. *BMC Genomics*, 6:100, 2005.

[15] X. H. Chen, A. P. Dempster, and J.S. Liu. Weighted finite population sampling to maximize entropy. *Biometrika*, 81:457–469, 1994.

[16] J. Cheng and R. Greiner. Comparing Bayesian network classifiers. Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, Stockholm, 1999. Morgan Kaufmann.

[17] C.S. Chow and D.F. Wirth. Linker scanning mutagenesis of the Plasmodium gallinaceum sexual stage specific gene pgs28 reveals a novel downstream cis-control element. *Molecular & Biochemical Parasitology*, 129:199–208, 2003.

[18] S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P.O. Brown, and I. Herskowitz. The transcriptional program of sporulation in budding yeast. *Science*, 282:699–705, 1998.

[19] H.M. Connolly, H.V. Schaff, C.J. Mullany, J. Rubin, M.D. Abel, and P.A. Pellikka. Surgical management of left-sided carcinoid heart disease. *Circulation*, 104:I36–I40, 2001.

[20] G.F. Cooper. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42:393–405, 1990.

[21] G.E. Crooks, G. Hon, J.-M. Chandonia, and S.E. Brenner. WebLogo: A Sequence Logo Generator. *Genome Research*, 14:1188–1190, 2004.

[22] J. Darroch. On the distribution of the number of successes in independent trials. *The Annals of Mathematical Statistics*, 35:1317–1321, 1964.

[23] F.T. de Dombal, D. Leaper, J. Staniland, J. Horrocks, and A. McCann. Computer aided diagnosis of acute abdominal pain. *British Medical Journal*, 2:9–13, 1972.

[24] K.J. Dechering, A.M. Kaan, W. Mbacham, D.F. Wirth, W. Eling, R.N.H. Konings, and H.G. Stunnenberg. Isolation and functional characterization of two distinct sexual-stage-specific promoters of the human malaria parasite Plasmodium falciparum. *Molecular & Cellular Biology*, 19:967–978, 1999.

[25] R. Dechter. Bucket elimination: A unifying framework for probabilistic inference. In *Proceedings of the Twelthth Conference on Uncertainty in Artificial Intelligence*, pages 211–219, 1996.

[26] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39:1–38, 1977.

[27] F.J. Díez. Parameter adjustment in Bayes networks. The generalized noisy OR-gate. In *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence*, pages 99–105, 1993.

[28] E. Dimitriadou, S. Dolničar, and A. Weingessel. An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika*, 67:137–160, 2002.

[29] P. Domingos and M. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29(2): 103–130, 1997.

[30] A.W.P. Edwards. The meaning of binomial distribution. *Nature*, 186:1074, 1960.

[31] H.B. Enderton. *A Mathematical Introduction to Logic*. Academic Press, San Diego, 1972.

[32] W. Feller. *An Introduction to Probability Theory and Its Applications*. John Wiley, 1968.

[33] L. Florens, M.P. Washburn, J.D. Raine, R.M. Anthony, M. Grainger, J.D. Haynes, J.K. Moch, N. Muster, J.B. Sacci, D.L. Tabb, A.A. Witney, D. Wolters, Y. Wu, M.J. Gardner, A.A. Holder, R.E. Sinden, J.R. Yates, and D.J. Carucci. A proteomic view of the Plasmodium falciparum life cycle. *Nature*, 419:520–526, 2002.

[34] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29(2):131–163, 1997.

[35] J. Gallup and J. Sachs. The economic burden of malaria. *The American Journal of Tropical Medicine and Hygiene*, 64:85–96, 2001.

[36] M. J. Gardner, N. Hall, E. Fung, O. White, M. Berriman, R. W. Hyman, J. M. Carlton, A. Pain, K. E. Nelson, S. Bowman, I. T. Paulsen, K. James, J. A. Eisen, K. Rutherford, S. L. Salzberg, A. Craig, S. Kyes, M. S. Chan, V. Nene, S. J. Shallom, B. Suh, J. Peterson, S. Angiuoli, M. Pertea, J. Allen, J. Selengut, D. Haft, M. W. Mather, A. B. Vaidya, D. M. Martin, A. H. Fairlamb, M. J. Fraunholz, D. S. Roos, S. A. Ralph, G. I. McFadden, L. M. Cummings, G. M. Subramanian, C. Mungall, J. C. Venter, D. J. Carucci, S. L. Hoffman, C. Newbold, R. W. Davis, C. M. Fraser, and B. Barrell. Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature*, 419:498–511, 2002.

[37] J.L. Gastwirth. A probability model of a pyramid scheme. *The American Statistician*, 31:79–82, 1977.

[38] M.H. Gelb. Drug discovery for malaria: a very challenging and timely endeavor. *Current Opinion in Chemical Biology*, 11:440–445, 2007.

[39] D. GuhaThakurta and G. Stormo. Identifying target sites for cooperatively binding factors. *Bioinformatics*, 17:608–621, 2001.

[40] A.M. Gunasekera, A. Myrick, K.T. Militello, J.S. Sims, C.K. Dong, T. Gierahn, K. Le Roch, E. Winzeler, and D.F. Wirth. Regulatory

motifs uncovered among gene expression clusters in Plasmodium falciparum. *Molecular & Biochemical Parasitology*, 153:19–30, 2007.

[41] C.T. Harbison, D.B. Gordon, T.I. Lee, N.J. Rinaldi, K.D. MacIsaac, T.W. Danford, N.M. Hannett, J.B. Tagne, D.B. Reynolds, J. Yoo, E.G. Jennings, J. Zeitlinger, D.K. Pokholok, M. Kellis, P.A. Rolfe, K.T. Takusagawa, E.S. Lander, D.K. Gifford, E. Fraenkel, and R.A. Young. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431:99–104, 2004.

[42] D. Heckerman. Causal independence for knowledge acquisition and inference. In *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence*, pages 122–127, 1993.

[43] D. Heckerman and J.S. Breese. A new look at causal independence. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pages 286–292, 1994.

[44] J. Hemingway and H. Ranson. Insecticide resistance in insect vectors of human disease. *Annual Review of Entomology*, 45:369–389, 2000.

[45] Max Henrion. Some practical issues in constructing belief networks. In *Proceedings of the Third Annual Conference on Uncertainty in Artificial Intelligence*, pages 161–174, 1987.

[46] K. Higo, Y. Ugawa, M. Iwamoto, and T. Korenaga. Plant cis-acting regulatory DNA elements (PLACE) database. *Nucleic Acids Research*, 27:297–300, 1999.

[47] W. Hoeffding. On the distribution of the number of successes in independent trials. *The Annals of Mathematical Statistics*, 27:713–721, 1956.

[48] P. Horrocks and M. Lanzer. Mutational analysis identifies a five base pair cis-acting sequence essential for GBP130 promoter activity in Plasmodium falciparum. *Molecular & Biochemical Parasitology*, 99:77–87, 1999.

[49] S. Howard. Discussion on Professor Cox's paper. *Journal of the Royal Statistical Society*, 34:210–211, 1972. Series B.

[50] L.J. Hubert and J.R. Levin. A general statistical framework for accessing categorical clustering in free recall. *Psychological Bulletin*, 83:1072–1082, 1976.

[51] J.D. Hughes, P.W. Estep, Tavazoie S., and G.M. Church. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *Journal of Molecular Biology*, 296:1205–1214, 2000.

[52] T.R. Hvidsten, B. Wilczyński, A. Kryshtafovych, J. Tiuryn, J. Komorowski, and Fidelis K. Discovering regulatory binding-site modules using rule-based learning. *Genome Research*, 15:856–866, 2005.

[53] S. Janson. Large deviation inequalities for sums of indicator variables. Technical report, Uppsala University, 1994.

[54] F.V. Jensen, S.L. Lauritzen, and K.G. Olesen. Bayesian updating in causal probabilistic networks by local computations. *Computational Statistics Quaterly*, 4:269–282, 1990.

[55] K. Jogdeo and S.M. Samuels. Monotone convergence of binomial probabilities and a generalization of Ramanujan's equation. *The Annals of Mathematical Statistics*, 39:1191–1195, 1968.

[56] M. Jordan, Z. Ghaharamani, T. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

[57] H.J. Kappen and J.P. Neijt. Promedas, a probabilistic decision support system for medical diagnosis. Technical report, SNN - UMCU, 2002.

[58] G. Karčiauskas. Text categorization using hierarchical Bayesian network classifiers. Master's thesis, Aalborg University, 2002.

[59] A.E. Kazakov, M.J. Cipriano, P.S. Novichkov, S. Minovitsky, D.V. Vinogradov, A. Arkin, A.A. Mironov, M.S. Gelfand, and I. Dubchak. RegTransBase - a database of regulatory sequences and interactions in a wide range of prokaryotic genomes. *Nucleic Acids Research*, 35: 407–412, 2007.

[60] S. Keleş, M. van der Laan, and M.B. Eisen. Identification of regulatory elements using a feature selection method. *Bioinformatics*, 18: 1167–1175, 2002.

[61] J.M.M. Kelly, L. McRobert, and D.A.A. Baker. Evidence on the chromosomal location of centromeric DNA in Plasmodium falciparum from etoposide-mediated topoisomerase-II cleavage. *Proceedings of the National Academy of Sciences of the United States of America*.

[62] R. Kline. *Principles and Practice of Structural Equation Modeling*. Guilford, New York, NY, 1998.

[63] R. Kohavi and D.H. Wolpert. Bias plus variance decomposition for zero-one loss functions. In L. Saitta, editor, *Machine Learning: Proceedings of the Thirteenth International Conference*, pages 275–283. Morgan Kaufmann, San Mateo, CA, 1996.

[64] M. Krugliak, J. Zhang, and H. Ginsburg. Intraerythrocytic Plasmodium falciparum utilizes only a fraction of the amino acids derived from the digestion of host cell cytosol for the biosynthesis of its proteins. *Molecular & Biochemical Parasitology*, 119:249–256, 2002.

[65] C. Lacave and F. Díez. A review of explanation methods for Bayesian networks. *Knowledge Engineering Review*, 17(2):107–127, 2002.

[66] M. Lanzer, D. de Bruin, and J.V. Ravetch. A sequence element associated with the Plasmodium falciparum KAHRP gene is the site of developmentally regulated protein-DNA interactions. *Nuclear Acids Research*, 20:3051–3056, 1992.

[67] S.L. Lauritzen. The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis*, 19: 191–201, 1995.

[68] S.L. Lauritzen and D.J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society*, 50:157–224, 1988.

[69] L. Le Cam. An approximation theorem for the Poisson binomial distribution. *Pacific Journal of Mathematics*, 10:1181–1197, 1960.

[70] R.S. Ledley and L.B. Lusted. Reasoning foundations of medical diagnosis: Symbolic logic, probability, and value theory aid our understanding of how physicians reason. *Science*, 130(3366):9–21, July 3, 1959 1959.

[71] P.J.F. Lucas. Bayesian network modelling through qualitative patterns. *Artificial Intelligence*, 163:233–263, 2005.

[72] P.J.F. Lucas, H. Boot, and B. Taal. Computer-based decision support in management of primary gastric non-Hodgkin lymphoma. *Methods of Information in Medicine*, 37:206–219, 1998.

[73] K.D. MacIsaac, T. Wang, D.B. Gordon, D.K. Gifford, G. Stormo, and E. Fraenkel. An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics*, 7:113, 2006.

[74] S. Mahony and P.V. Benos. STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Research*, 35:253–258, 2007.

[75] V. Matys, E. Kel-Margoulis, O.V.and Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, M. Chekmenev, D.and Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A.E. Kel, and E. Wingender. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research*, 34:108–110, 2006.

[76] C. Meek and D. Heckerman. Structure and parameter learning for causal independence and causal interaction models. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, pages 366–375, 1997.

[77] K.T. Militello, M. Dodge, L. Benthke, and D.F. Wirth. Identification of regulatory elements in the Plasmodium falciparum genome. *Molecular & Biochemical Parasitology*, 134:75–88, 2004.

[78] I.M. Modlin, M.D. Shapiro, and M. Kidd. Carcinoid tumors and fibrosis: An association with no explanation. *The American Journal of Gastroenterology*, 99(12):2466–2478, 2004.

[79] S. Musunuru, J.E. Carpenter, R.S. Sippel, M. Kunnimalaiyaan, and H. Chen. A mouse model of carcinoid syndrome and heart disease. *The Journal of Surgical Research*, 126:102–105, 2005.

[80] R.M. Neal. Probabilistic inference using Markov chain Monte Carlo methods. Technical report, University of Toronto, 1993. CRG-TR-93-1.

[81] R.M. Neal and G.E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In M.I. Jordan, editor, *Learning in Graphical Models*. MIT Press, Cambridge, 1998.

[82] F.R.E. Nobels, D.J. Kwekkeboom, R. Bouillon, and S.W.J. Lamberts. Chromogranin a: its clinical value as marker of neuroendocrine tumours. *European Journal of Clinical Investigation*, 28: 431–440, 1998.

[83] K.G. Olesen, U. Kjæ rulff, F. Jensen, B. Falck, S. Andreassen, and S.K. Andersen. A munin network for the median nerve - a case study on loops. *Applied Artificial Intelligence*, 3:384–403, 1989.

[84] M. Osta, L. Gannoun-Zaki, S. Bonnefoy, C. Roy, and H.J. Vial. A 24 bp cis-acting element essential for the transcriptional activity of Plasmodium falciparum CDP-diacylglycerol synthase gene promoter. *Molecular & Biochemical Parasitology*, 121:87–98, 2002.

[85] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kauffman Publishers, 1988.

[86] O.E. Percus and J.K. Percus. Probability bounds on the sum of independent nonidentically distributed binomial random variables. *SIAM Journal on Applied Mathematics*, 45:621–640, 1985.

[87] Y. Pilpel, P. Sudarsanam, and G.M. Church. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genetics*, 29:153–159, 2001.

[88] J. Pitman. Probabilistic bounds on the coefficients of polynomials with only real zeros. *Journal of Combinatorial Theory*, 77:279–303, 1997. Series A.

[89] M.E. Porter. Positive and negative effects of deletions and mutations within the 5' flanking sequences of Plasmodium falciparum DNA polymerase δ. *Molecular & Biochemical Parasitology*, 122:9–19, 2002.

[90] B. Roos. Binomial approximation to the poisson binomial distribution: the krawtchouk expansion. *Theory of Probability and Its Applications*, 45:258–272, 2001.

[91] D. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York, NY, 1987.

[92] O. K. Ruvalcaba-Salazar, M. del Carmen Ramírez-Estudillo, D. Montiel-Condado, F. Recillas-Targa, M. Vargas, and R. Hernández-Rivas. Recombinant and native Plasmodium falciparum TATA-binding-protein binds to a specific TATA box element in promoter regions. *Molecular & Biochemical Parasitology*, 140:183–196, 2005.

[93] J.D. Sachs and P. Malaney. The economic and social burden of malaria. *Nature*, 415:680–685, 2002.

[94] M. Sahami. Learning limited dependence bayesian classifiers. Conference on Knowledge Discovery in Databases, Portland, OR, 1996.

[95] T. Sakata and E.A. Winzeler. Genomics, systems biology and drug development for infectious diseases. *Molecular BioSystems*, 3:841–848, 2007.

[96] S. Salzberg. On comparing classifiers: pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, 1: 317–327, 1997.

[97] A. Sandelin, W. Alkema, P. Engstrom, W.W. Wasserman, and B. Lenhard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, 32:91–94, 2004.

[98] A.A. Schäffer, L. Aravind, T.L. Madden, S. Shavirin, J.L. Spouge, Y.I. Wolf, E.V. Koonin, and S.F. Altschul. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Research*, 29:2994–3005, 2001.

[99] E. Segal, R. Yelensky, and D. Koller. Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics*, 19:1273–1282, 2003.

[100] R.D. Shachter and M.A. Peot. Simulation approaches to general probabilistic inference on belief networks. In *Proceedings of the Fifth Conference on Uncertainty in Artificial Intelligence*, pages 221–231, 1989.

[101] G. Shafer and P. Shenoy. Probability propagation. *Annals of Mathematics and Artificial Intelligence*, 2:327–352, 1990.

[102] J.L. Shock, K.F. Fischer, and J.L. DeRisi. Whole-genome analysis of mRNA decay in Plasmodium falciparum reveals a global lengthening of mRNA half-life during the intra-erythrocytic development cycle. *Genome Biology*, 8:R134, 2007.

[103] M.A. Shwe, B. Middleton, D.E. Heckerman, M. Henrion, E.J. Horvitz, H.P. Lehmann, and G.F. Cooper. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base, I – The probabilistic model and inference algorithms. *Methods of Information in Medicine*, 30:241–255, 1991.

[104] D.J. Spiegelhalter and R.P. Knill-Jones. Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology. *Journal of the Royal Statistical Society. Series A (General)*, 147(1):35–77, 1984.

[105] S. Srinivas. A generalization of the noisy-or model. In *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence*, pages 208–215, 1993.

[106] N.O. Steffens, C. Galuschka, M. Schindler, L. Bülow, and R. Hehl. AthaMap: an online resource for in silico transcription factor binding sites in the Arabidopsis thaliana genome. *Nucleic Acids Research*, 32:368–372, 2004.

[107] S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho, and G.M. Church. Systematic determination of genetic network architecture. *Nature Genetics*, 22, 1999.

[108] R. Teach and E. Shortliffe. An analysis of physician attitudes regarding computerbased clinical consultation systems. *Computers and Biomedical Research*, 14:542–558, 1981.

[109] K.K. Tetteh and S.D. Polley. Progress and challenges towards the development of malaria vaccines. *BioDrugs*, 21:357–373, 2007.

[110] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R.B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17:520–525, 2001.

[111] P. van Beek. An application of Fourier methods to the problem of sharpening the Berry-Essen inequality. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 23:187–196, 1972.

[112] M.A.J. van Gerven and P.J.F. Lucas. Using background knowledge to construct Bayesian classifiers for data-poor domains. pages 269–282, London, UK, 2004. Twenty-fourth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence.

[113] M.A.J. van Gerven, P.J.F. Lucas, and T.P. Van der Weide. A generic qualitative characterization of independence of causal influence. *International Journal of Approximate Reasoning*, 48:214–236, 2007.

[114] V. Van Noort and M.A. Huynen. Combinatorial gene regulation in Plasmodium falciparum. *Trends in Genetics*, 22:73–78, 2006.

[115] C. van Rijsbergen. *Information Retrieval*. Butterworths, London, UK, 1979.

[116] S. Visscher, P.J.F. Lucas, M. Bonten, and K. Schurink. Improving the therapeutic performance of a medical Bayesian network using noisy threshold models. volume 3745 of *Proceedings of ISBMDA 2005, the 6th International Symposium on Biological and Medical Data Analysis*, pages 161–172, 2005.

[117] J. Vomlel. Exploiting functional dependence in bayesian network inference. In *Eighteenth Conference on Uncertainty in Artificial Intelligence*, page 528535, San Francisco, CA, 2002.

[118] J. Vomlel. Noisy-or classifier. *International Journal of Intelligent Systems*, 21:381–398, 2005.

[119] A. Wagner. Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics*, 15:776–784, 1999.

[120] I. Wegener. *The Complexity of Boolean Functions*. John Wiley & Sons, New York, 1987.

[121] T. Werner. Models for prediction and recognition of eukaryotic promoters. *Mammalian Genome*, 10:168–175, 1999.

[122] I.H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2005.

[123] X. Xie, J. Lu, E.J. Kulbokas, T.R. Golub, V. Mootha, K. Lindblad-Toh, E.S. Lander, and M. Kellis. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, 434:338–345, 2005.

[124] J.A. Young, J.R. Johnson, C. Benner, S.F. Yan, K. Chen, K.G. Le Roch, Y. Zhou, and E.A. Winzeler. In silico discovery of transcription regulatory elements in Plasmodium falciparum. *BMC Genomics*, 9:70, 2008.

[125] Y. Yuan, L. Guo, L. Shen, and J.S. Liu. Predicting gene expression from sequence: a reexamination. *PLoS Computational Biology*, 3: e243, 2007.

[126] A. Zagorecki, M. Voortman, and M. Druzdzel. Decomposing local probability distributions in Bayesian networks for improved inference and parameter learning. In *Proceedings of the Nineteenth International Florida Artificial Intelligence Research Society Conference*, pages 860–865, 2006.

[127] N.L. Zhang and D. Poole. A simple approach to Bayesian network computations. In *Proceedings of the Tenth Canadian Conference on Artificial Intelligence*, pages 171–178, 1994.

[128] N.L. Zhang and D. Poole. Exploiting causal independence in Bayesian networks inference. *Journal of Artificial Intelligence Research*, 5:301–328, 1996.

[129] J.M. Zuetenhorst and B.G. Taal. Metastatic carcinoid tumors: A clinical review. *Oncologist*, 10(2):123–131, February 1, 2005 2005.

[130] J.M. Zuetenhorst, J.M. Bonfrer, C.M. Korse, R. Bakker, H. van Tinteren, and B.G. Taal. Carcinoid heart disease. *Cancer*, 97(7):1609–1615, 2003.

Summary

This thesis studies in detail symmetric causal independence models, which are a type of Bayesian network.

A *Bayesian network* provides a graphical representation of a probability distribution over a set of random variables; it consists of nodes that represent variables and arcs that encode conditional independencies between the variables. Bayesian networks have a number of attractive properties. Firstly, they provide a framework for deriving efficient algorithms for *probabilistic inference*, the task of computing conditional probabilities of the values of some of the nodes given the values of other nodes. Secondly, the models provide an intuitive representation of domain knowledge. Thirdly, Bayesian networks allow combining expert knowledge and statistical data. Finally, Bayesian networks are well suited to dealing with incomplete data.

The definition of a Bayesian network does not constrain how a variable depends on its parents (variables that have outgoing arcs into the variable). However, the number of conditional probabilities for a variable grows exponentially with the number of its parents, making the tasks of specifying the conditional probabilities and probabilistic inference in richly-connected Bayesian networks difficult or even intractable. Therefore, researchers proposed a number of ways of economically specifying the conditional probability of a variable with many parents.

Causal independence is a popular way to constrain the conditional probability tables for binary variables. The global structure of a causal independence model is shown in Figure 1; it expresses the idea that causes $C_1, \ldots, C_n$ influence a given common effect E through hidden variables $H_1, \ldots, H_n$ and the interaction function f. The hidden variable H_i is considered to be a contribution of the cause variable C_i to the common effect E, and absent causes do not contribute to the effect. The function

f defines the way in which the hidden effects $H_1, \ldots, H_n$ and, indirectly, also the causes $C_1, \ldots, C_n$ interact to yield the final effect E. Number of parameters in causal independence models is linear in the number of causes.

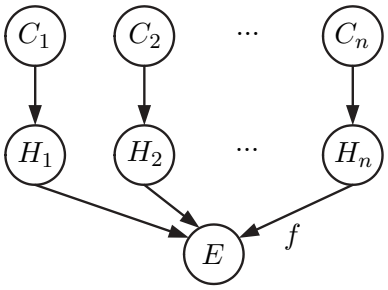


Figure 1: Causal independence model.

While many real-world Bayesian networks incorporate causal independence assumptions, only two interaction functions, the logical OR function and the logical AND function, are used in practice. The underlying assumption of the resulting models, known as noisy-OR and noisy-AND models, is that the presence of either at least one cause or all causes at the same time gives rise to the effect. In this thesis, we study a broader class of causal independence models, causal independence models based on the symmetric Boolean functions. We pay special attention to causal independence models based on a useful symmetric Boolean function, the Boolean threshold function τ_k, which checks whether there are at least k arguments whose value is true. Causal independence models based on the symmetric Boolean function and the Boolean threshold function are further referred to as symmetric causal independence models and noisy threshold models, respectively.

In Chapter 2, we investigate the problem of probabilistic inference in symmetric causal independence models, with a special focus on the noisy threshold models. We establish a connection between the conditional probability distribution of the effect variable in these models and the Poisson binomial distribution. We investigate how the properties of the Poisson binomial distribution can be used for computationally efficient exact and approximate inference in symmetric causal independence models. We also compare the efficiency of the computational schemes developed with the efficiency of standard inference techniques.

Learning a Bayesian network from data includes two tasks: learning the

structure and learning the parameters. In case of symmetric causal independence models, we deal with models whose structure is fixed; therefore, in order to learn symmetric causal independence models, we need to learn only their parameters. The problem of learning the parameters in symmetric causal independence models is studied in Chapter 3. We present a computationally efficient expectation-maximization (EM) algorithm to learn parameters in symmetric causal independence models, where the computational scheme of the Poisson binomial distribution is used to compute the conditional probabilities in the E-step. We study computational complexity and convergence of the developed algorithm. To assess the practical usefulness of symmetric causal independence models, we use two data sets that are different in their causal interpretation and size. The non-Hodgkin lymphoma data set consists of factors that influence the result of the treatment, and, for this reason, the models learned from this data set can be argued to follow the causal interpretation. The second data set consisting of Reuters news stories allowed us to test the EM algorithm on large symmetric causal independence models as the number of cause variables for some document classes is in the hundreds. The models learned from these data sets were applied to a classification task and shown to perform competitively with state-of-the-art classifiers.

Chapter 4 presents the application of the noisy threshold model to predict whether a patient with carcinoid syndrome will develop a carcinoid heart disease. Carcinoid heart disease is the most dangerous complication of carcinoid syndrome as it occurs in over 65 percent of patients with carcinoid syndrome and is a major source of morbidity and mortality for patients with this syndrome. Given that so many carcinoid patients die of carcinoid heart disease, it is important to divide patients that are admitted to the clinic into patients that are likely to develop a severe form of carcinoid heart disease, and those that do not develop this severe form. In this way, patients that are at risk can be given more aggressive treatment in order to reduce the probability of the development of carcinoid heart disease. We use data of fifty-four carcinoid patients, of which twenty-two patients developed carcinoid heart disease. The noisy threshold model performed favorably to four state-of-the-art classification algorithms, and equally well as a decision-rule that was formulated by the physician.

In Chapter 5, we use noisy threshold models to learn more about one of the processes fundamental to *Plasmodium* biology, transcriptional gene regulation. A parasite *Plasmodium* causes malaria, an infectious disease, which infects between 300 and 500 million people every year and accounts

for more than one million deaths annually. A thorough understanding of gene regulation in this organism is important for developing a better vaccine and identifying novel drug targets to fight this lethal disease. We use noisy threshold models to identify regulatory sequence elements explaining gene's membership to a gene expression cluster. Differently from other bioinformatics approaches, our method is able to model the logic behind gene regulation and to incorporate uncertainty about the functionality of putative regulatory sequence elements. We apply our technique to *Plasmodium falciparum*, the most virulant species of the four species of *Plasmodia* affecting humans. Our analysis finds thirty-nine putative regulatory sequence elements involved in gene regulation during the intraerythrocytic developmental cycle. Furthermore, we find no evidence that additional information about regulatory sequence elements improves prediction of gene expression. Finally, we provide a list of ten potential transcription factors.

Samenvatting

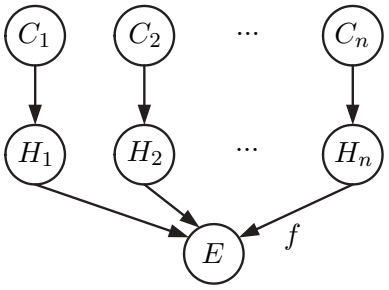
Dit proefschrift beschouwt een specifiek type van Bayesiaanse netwerken, namelijk symmetrische causale onafhankelijkheidsmodellen.

Een *Bayesiaans netwerk* is een grafische representatie van de kansverdeling van een set variabelen met een random verdeling. Het bestaat uit een aantal knopen die variabelen representeren en pijlen die de voorwaardelijke onafhankelijkheid tussen de variabelen weergeven. Bayesiaanse netwerken hebben een aantal voordelige eigenschappen. Ten eerste bieden ze een raamwerk dat gebruikt kan worden om efficiënte algoritmes te ontwikkelen voor de probabilistische inferentie, de berekening van de conditionele waarschijnlijkheid van de waardes van bepaalde knopen gegeven de waardes van andere knopen. Ten tweede bieden deze netwerken een intuïtieve representatie van domeinkennis. Ten derde kunnen Bayesiaanse netwerken gebruikt worden om de kennis van experts te combineren met statistische data. Als laatste zijn Bayesiaanse netwerken geschikt om met incomplete data om te gaan.

De definitie van een Bayesiaans netwerk is niet beperkend voor hoe een variabele afhangt van haar ouders (variabelen met uitgaande pijlen naar deze variabele). Het aantal conditionele waarschijnlijkheden van een variabele neemt echter exponentieel toe met het aantal ouders, zodat het specificeren van de conditionele waarschijnlijkheid en de probabilistische inferentie voor Bayesiaanse netwerken met veel connecties moeilijk of zelfs onmogelijk is. Daarom zijn er in de literatuur een aantal manieren voorgesteld om de conditionele waarschijnlijkheid van een variabele met veel ouders op een meer economische manier te specificeren.

Causale onafhankelijkheid is een handige manier om de conditionele waarschijnlijkheidstabellen voor binaire variabelen te beperken. De globale structuur van een causaal onafhankelijkheidsmodel is weergegeven in Fi-

guur 1. De achterliggende gedachte is dat de oorzaken $C_1, \ldots, C_n$ het gezamenlijke effect E beïnvloeden via de verborgen variabelen $H_1, \ldots, H_n$ en de interactiefunctie f. De verborgen variabele H_i wordt beschouwd als de contributie van de causale variabele C_i op het gezamenlijke effect E. Niet aanwezige oorzaken dragen ook niet bij aan dit effect. De functie f definieert de manier waarop de verborgen effecten $H_1, \ldots, H_n$ en indirect ook de oorzaken $C_1, \ldots, C_n$ met elkaar samenwerken om het uiteindelijke effect E te bewerkstelligen. Het aantal parameters in causale onafhankelijkheidsmodellen neemt lineair toe met het aantal oorzaken.



Figuur 1: Causaal onafhankelijkheidsmodel.

Ondanks dat veel Bayesiaanse netwerken gebruik maken van causale onafhankelijkheidsaannames, worden er slechts twee interactiefuncties, namelijk de logische OR en de logische AND functie in de praktijk gebruikt. De aannames die ten grondslag liggen aan deze modellen, ook wel bekend als noisy-OR en noisy-AND modellen, is dat de aanwezigheid van óf ten minste één oorzaak, of alle oorzaken tegelijkertijd het gezamenlijke effect veroorzaken. In dit proefschrift bestuderen we een grotere klasse van causale onafhankelijkheidsmodellen, namelijk causale onafhankelijkheidsmodellen die gebaseerd zijn op symmetrische Booleaanse functies. De focus ligt hierbij voornamelijk op causale onafhankelijkheidsmodellen die gebaseerd zijn op de Booleaanse drempelfunctie τ_k, die gebruikt kan worden om te testen of er tenminste k argumenten de waarde 'true' hebben. Causale onafhankelijkheidsmodellen die gebaseerd zijn op de symmetrische Booleaanse functie en de Booleaanse drempelfunctie worden vanaf nu respectievelijk symmetrische causale onafhankelijkheidsmodellen en noisy threshold modellen genoemd.

In hoofdstuk 2 bekijken we het probleem van probabilistische inferentie in symmetrische causale onafhankelijkheidsmodellen met de nadruk op de noisy threshold modellen. We leggen een verband tussen de voorwaardelijke kansverdeling van de effectvariabele in deze modellen en de Poisson

binomiale verdeling. We onderzoeken hoe de eigenschappen van de Poisson binomiale verdeling gebruikt kunnen worden voor efficiënte exacte en benaderende inferentie in symmetrische causale onafhankelijkheidsmodellen. Verder vergelijken we de efficiëntie van de ontwikkelde berekeningsmethoden met de efficiëntie van standaard inferentietechnieken.

Het leren van Bayesiaanse netwerken op basis van data bestaat uit twee taken: het leren van de structuur en het leren van de parameters. In het geval van symmetrische causale onafhankelijkheidsmodellen hebben we te maken met modellen waarvan de structuur vast staat. Daarom hoeven alleen de parameters geleerd te worden om een symmetrisch causaal onafhankelijkheidsmodel te leren. Het probleem van het leren van de parameters in deze modellen wordt behandeld in hoofdstuk 3. We presenteren een rekenkundig efficiënt expectation-maximization (EM) algoritme om de parameters te leren, waarbij het rekenkundige model van de Poisson binomiale verdeling gebruikt wordt in de E-stap. Ook bestuderen we de rekenkundige complexiteit en de convergentie van het ontwikkelde algoritme. Om de practische waarde van symmetrische causale onafhankelijkheidsmodellen te bepalen gebruiken we twee datasets die verschillen in de causale interpretatie en in grootte. De dataset horende bij het non-Hodgkin lymfoom bestaat uit de factoren die het resultaat van de behandeling beïnvloeden en daarom kan verondersteld worden dat de op basis van deze set geleerde modellen de causale interpretatie volgen. De tweede dataset bestaat uit Reuters nieuwsartikelen en stelt ons in staat om het EM algoritme te testen op grote symmetrische onafhankelijkheidsmodellen aangezien het aantal causale variabelen voor een aantal documentklassen in de honderden loopt. De op basis van deze twee sets geleerde modellen zijn toegepast op een classificeringstaak en hieruit blijkt dat ze competitief zijn ten opzichte van state-of-the-art classificatie algoritmes.

In hoofdstuk 4 passen we het noisy threshold model toe om te voorspellen of een patiënt met het carcinoid syndroom ook carcinoid hartziekte zal ontwikkelen. Carcinoid hartziekte is de meest gevaarlijke complicatie van het carcinoid syndroom omdat het bij 65 procent van de patiënten met het carcinoid syndroom voorkomt en het één van de hoofdoorzaken is van de sterfte van patiënten met dit syndroom. Aangezien er een groot aantal hartpatiënten sterft aan carcinoid hartziekte is het belangrijk om bij de opname van patiënten in de kliniek een scheiding te maken tussen patiënten die waarschijnlijk een ernstige vorm van carcinoid hartziekte ontwikkelen en degenen die dit waarschijnlijk niet zullen ontwikkelen. Op deze manier kunnen er voor de risicogroep aggresievere behandelmethoden gebruikt

worden om de kans te verkleinen om carcinoid hartziekte te ontwikkelen. We gebruiken data van 54 hartpatiënten waarvan er 22 carcinoid hartziekte hebben ontwikkeld. Het noisy threshold model presteert beter dan vier state-of-the-art classificatie algoritmes en evengoed als het beslissingsschema dat is ontwikkeld door de arts.

In hoofdstuk 5 gebruiken we noisy threshold modellen om meer inzicht te krijgen in één van de fundamentele processen van de *Plasmodium* biologie, namelijk de transcriptionele regulatie van de genexpressie bij de malariaparasiet. De *Plasmodium* parasiet veroorzaakt jaarlijks de infectieziekte malaria bij 300 tot 500 miljoen personen en is verantwoordelijk voor de dood van meer dan 1 miljoen mensen per jaar. Een goed begrip van de regulatie van de genexpressie van dit organisme is van groot belang voor het ontwikkelen van betere vaccinaties en het bepalen van nieuwe doelen voor medicijnen om de dodelijke ziekte tegen te gaan. We gebruiken noisy threshold modellen om de regulatorische elementen te identificeren die kunnen verklaren waarom een bepaald gen tot een genexpressiecluster behoort. In tegenstelling tot andere aanpakken in de bioinformatica kan onze methode gebruikt worden om de logica achter de regulatie van de genexpressie te modelleren en om onzekerheid over de functionaliteit van de vermoedelijke regulatorische elementen mee te nemen. We passen onze aanpak toe op *Plasmodium falciparum*, de meest dodelijke variant van de vier soorten *Plasmodia* die schadelijk zijn voor de mens. Onze analyse vindt negenendertig regulatorische elementen die een rol spelen bij de regulatie van de genexpressie tijdens de intraerythrociete ontwikkelcyclus. Verder vinden we geen bewijs dat extra informatie over regulatorische elementen de voorspelling van de genexpressie verbetert. Als laatste presenteren we een lijst van tien mogelijke transcriptiefactoren.

Santrauka

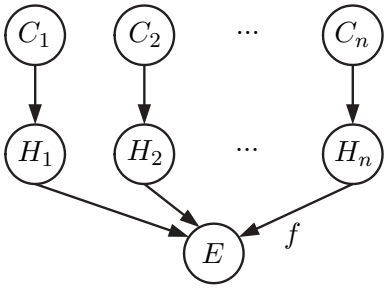
Šioje disertacijoje detaliai nagrinėjama viena iš Bajeso tinklų klasių - simetriniai priežasčių nepriklausomybės modeliai (angl. symmetric causal independence models).

Bajeso tinklas - tai grafinis atsitiktinių dydžių aibės tikimybinio skirstinio atvaizdavimas. Jis sudarytas iš mazgų, kurie atvaizduoja atsitiktinius dydžius, ir rodyklių, kurių nebuvimas ženklina sąlygines nepriklausomybes tarp atsitiktinių dydžių. Bajeso tinklai yra patrauklūs dėl keleto savybių. Pirma, jų struktūra suteikia galimybę sukurti efektyvius *tikimybinės išvados* (angl. probabilistic inference) algoritmus, t.y. algoritmus, randančius grupės mazgų reikšmių sąlygines tikimybes žinant kitų mazgų reikšmes. Antra, šie modeliai suteikia galimybę intuityviai atvaizduoti srities žinias. Trečia, Bajeso tinklai leidžia suderinti specialistų žinias ir statistinius duomenis. Galiausiai, Bajeso tinklai puikiai tinka tais atvejais, kai duomenys yra nepilni.

Bajeso tinklo apibrėžimas neriboja atsitiktinio dydžio priklausomybės nuo jo tėvų (atsitiktinių dydžių, kuriuos jungia išeinančios rodyklės su šiuo atsitiktiniu dydžiu). Tačiau kintamojo sąlyginių tikimybių skaičius auga eksponentiškai su kintamojo tėvų skaičiumi. To pasekoje sąlyginių tikimybių nustatymas ir tikimybinė išvada, turint gausiai sujungtus Bajeso tinklus, tampa sudėtingomis ar netgi neįveikiamomis užduotimis. Todėl tyrinėtojai pasiūlė daugybę būdų ekonomiškai apibrėžti atsitiktinio dydžio, turinčio daugybę tėvų, sąlygines tikimybes.

Priežasčių nepriklausomybė - tai populiarus būdas suvaržyti atsitiktinių binarinių dydžių sąlyginių tikimybių lenteles. Bendra priežasčių nepriklausomybės modelio struktūra (Paveikslas 1) parodo modelio esmę: priežastys $C_1, \ldots, C_n$ paveikia tam tikrą bendrą rezultatą E per paslėptuosius atsitiktinius dydžius $H_1, \ldots, H_n$ ir sąveikos funkciją f. Paslėptasis atsitiktinis

dydis H_i yra laikomas priežastinio atsitiktinio dydžio C_i indėliu į bendrą rezultatą E. Tuo tarpu nesamos priežastys neįtakoja teigiamo rezultato. Funkcija f apibrėžia kaip paslėptieji rezultatai $H_1, \dots H_n$, o taip pat netiesiogiai ir priežastys $C_1, \dots, C_n$ sąveikauja ir sukelia galutinį rezultatą E. Priežasčių nepriklausomybės modelių parametrų skaičius tiesiškai priklauso nuo priežasčių skaičiaus.



Paveikslas 1: Priežasčių nepriklausomybės modelis.

Nors ir daugybė praktinių Bajeso tinklų įtraukia priežasčių nepriklausomybės prielaidas, praktikoje yra naudojamos tik dvi sąveikos funkcijos: loginė ARBA funkcija ir loginė IR funkcija. Gauti modeliai, žinomi kaip triukšmingas-ARBA ir triukšmingas-IR modeliai (angl. noisy-OR and noisy-AND models), pagrįsti prielaidomis, kad bent vienos arba visų priežasčių vienu metu buvimas sąlygoja teigiamą rezultatą. Šioje disertacijoje mes tyrinėjame platesnę priežasčių nepriklausomybės modelių klasę - priežasčių nepriklausomybės modelius, pagrįstus simetrinėmis Bulio funkcijomis. Išskirtinis dėmesys skiriamas priežasčių nepriklausomybės modeliams, pagrįstiems naudinga simetrine Bulio funkcija - Bulio slenkstine funkcija τ_k. Ši funkcija tikrina ar yra bent k argumentai, kurių reikšmė yra teisinga. Priežasčių nepriklausomybės modeliai pagrįsti simetrine Bulio funkcija ir Bulio slenkstine funkcija toliau bus vadinami atitinkamai simetriniais priežasčių nepriklausomybės modeliais ir triukšmingais slenkstiniais modeliais (angl. noisy threshold models).

Antrame skyriuje tiriame tikimybinės išvados simetriniuose priežasčių nepriklausomybės modeliuose problemą, ypatingą dėmesį skirdami triukšmingiems slenkstiniams modeliams. Mes nustatome ryšį tarp rezultato sąlyginių tikimybių skirstinio šiuose modeliuose ir Puasono binominio skirstinio. Tiriama, kaip Puasono binominio skirstinio savybės gali būti panaudotos efektyvių tikslių ir apytikslių tikimybinės išvados simetriniuose priežasčių nepriklausomybės modeliuose algoritmų sukūrimui. Taip pat mes lyginame pasiūlytų skaičiavimo schemų efektyvumą su įprastų tikimybinės išvados

metodų efektyvumu.

Norint išmokti Bajeso tinklą iš duomenų, reikia išmokti jo struktūrą ir parametrus. Simetrinių priežasčių nepriklausomybės modelių atveju mes turime modelius, kurių struktūra nustatyta iš anksto. Todėl norint išmokti simetrinius priežasčių nepriklausomybės modelius, tereikia išmokti modelių parametrus. Simetrinių priežasčių nepriklausomybės modelių parametrų išmokimo problema nagrinėjama trečiame skyriuje. Čia pateikiamas efektyvus matematinės vilties maksimizavimo (angl. expectation-maximization; toliau - EM) algoritmas, išmokstantis simetrinių priežasčių nepriklausomybės modelių parametrus. EM algoritmas naudoja Puasono binominio skirstinio skaičiavimo schemą, galinčią apskaičiuoti rezultato sąlygines tikimybes E-žingsnyje. Mes tiriame šio algoritmo sudėtingumą ir konvergavimą. Simetrinių priežasčių nepriklausomybės modelių praktinei naudai įvertinti naudojame du duomenų rinkinius, vienas nuo kito besiskiriančius tiek priežastine interpretacija, tiek dydžiu. Ne-Hodžkino limfomos duomenų rinkinys sudarytas iš faktorių, kurie įtakoja gydymo rezultatą, todėl galima teigti, kad modeliai išmokti iš šio duomenų rinkinio sutinka su priežastine modelio interpretacija. Antrasis duomenų rinkinys, sudarytas iš Reuters naujienų, leido mums patikrinti EM algoritmą didelių simetrinių priežasčių nepriklausomybės modelių atveju: kai kurios dokumentų klasės turi šimtus priežastinių atsitiktinių dydžių. Modeliai išmokti iš šių duomenų rinkinių buvo pritaikyti klasifikavimo užduočiai ir parodyta, kad jų veiksmingumas nenusileidžia moderniausių klasifikatorių veiksmingumui.

Ketvirtame skyriuje pristatome triukšmingo slenkstinio modelio pritaikymą numatyti ar karcinoidiniu sindromu sergančiam ligoniui išsivystys karcinoidinė širdies liga. Karcinoidinė širdies liga yra pavojingiausia karcinoidinio sindromo komplikacija - nuo šios komplikacijos kenčia daugiau nei šešiasdešimt penki procentai ligonių. Be to, ši komplikacija yra pagrindinė karcinoidiniu sindromu sergančių ligonių sergamumo ir mirties priežastis. Žinant kiek daug karcinoidiniu sindromu sergančių ligonių miršta nuo karcinoidinės širdies ligos, svarbu išskirti hospitalizuotus ligonius į ligonius, kuriems tikriausiai pasireikš sunki karcinoidinės širdies ligos forma bei į ligonius, kuriems greičiausiai ši komplikacija nepasireikš. Taip ligoniams, priklausantiems rizikos grupei, gali būti paskirtas agresyvesnis gydymas ir sumažinta tikimybė, kad jiems išsivystys karcinoidinė širdies liga. Mes naudojame penkiasdešimt keturių karcinoidiniu sindromu sergančių ligonių duomenis, iš kurių dvidešimt dviems išsivystė karcinoidinė širdies liga. Triukšmingo slenkstinio modelio tikslumas buvo didesnis negu keturių moderniausų klasifikatorių ir beveik toks pats kaip ir gydytojo suformuluotos

sprendimo taisyklės.

Penktame skyriuje mes taikome triukšmingus slenkstinius modelius siekiant geriau suprasti vieną iš esminių *Plasmodium* biologinių procesų - genų raiškos reguliavimą transkripcijos lygyje. *Plasmodium* parazitas sukelia maliariją - infekcinę ligą, kuria kasmet suserga nuo 300 iki 500 milijonų žmonių, o nuo šios ligos miršta daugiau negu milijonas žmonių. Šio parazito genų raiškos reguliavimo nuodugnus supratimas svarbus norint sukurti vakciną ir nustatyti naujus vaistų taikinius, kurie padėtų kovoti su šia mirtina liga. Mes taikome triukšmingus slenkstinius modelius su tikslu nustatyti reguliuojančiuosius sekos elementus, paaiškinančius geno priklausomybę genų raiškos klasteriui. Skirtingai nuo kitų bioinformatikos metodų, mūsų metodas pajėgus tiek sumodeliuoti logiką, kuria pagrįstas genų raiškos reguliavimas, tiek įtraukti netikrumą dėl numanomų reguliuojančiųjų sekos elementų funkcionalumo. Metodą taikome mirtingiausiai iš keturių *Plasmodium* rūšių, užkrečiančių žmones - *Plasmodium falciparum*. Mūsų analizė atskleidė trisdešimt devynis spėjamus reguliuojančiuosius sekos elementus, atsakingus už genų raiškos reguliavimą eritrocitinio vystymosi ciklo metu. Mes neradome įrodymų, kad papildoma informacija apie reguliuojančiuosius sekos elementus padeda nuspėti genų raišką. Galiausiai, mes pateikiame dešimt tikėtinų transkripcijos faktorių.

Curriculum Vitae

I was born on April 28th, 1979 in Kaunas, Lithuania. In 1997, I got my upper secondary school diploma from Kaunas "Saulės" gymnasium, an upper secondary school with a specific focus on mathematics, computer science and physics. During the last year of gymnasium, I wanted to become a journalist. However, after working in a newspaper of city pupils, I dropped this plan and I chose probably the most popular study among pupils from the gymnasium - computer science. In 2001, I obtained my bachelor's degree in Computer Science at Vytautas Magnus University in Kaunas with the thesis entitled "Application of dynamic programing to find parallel connections between speech and the corresponding text". I continued my studies at Aalborg University, Denmark and obtained my master's degree in Knowledge and Data Engineering in 2003. My master's thesis was entitled "Model-based hierarchical clustering using Bayesian networks". Since September 2003, I have been working as a Ph.D. student at the Institute of Computing and Information Sciences at Radboud University Nijmegen. The work carried out during this period is presented in this thesis.

SIKS Dissertation Series

====
1998
====

1998-01 Johan van den Akker (CWI)
DEGAS - An Active, Temporal Database of Autonomous Objects

1998-02 Floris Wiesman (UM)
Information Retrieval by Graphically Browsing Meta-Information

1998-03 Ans Steuten (TUD)
A Contribution to the Linguistic Analysis of Business Conversations within the Language/Action Perspective

1998-04 Dennis Breuker (UM)
Memory versus Search in Games

1998-05 Eduard W.Oskamp (RUL)
Computerondersteuning bij Straftoemeting

====
1999
====

1999-01 Mark Sloof (VU)
Physiology of Quality Change Modelling;
Automated Modelling of Quality Change of Agricultural Products

1999-02 Rob Potharst (EUR)
Classification using Decision Trees and Neural Nets

1999-03 Don Beal (UM)
The Nature of Minimax Search

1999-04 Jacques Penders (UM)
The practical Art of Moving Physical Objects

1999-05 Aldo de Moor (KUB)
Empowering Communities: A Method for the Legitimate User-Driven Specification of Network Information Systems

1999-06 Niek J.E. Wijngaards (VU)
Re-design of Compositional Systems

1999-07 David Spelt (UT)
Verification Support for Object Database Design

1999-08 Jacques H.J. Lenting (UM)
Informed Gambling: Conception and Analysis of a Multi-Agent Mechanism for Discrete Reallocation

====
2000
====

2000-01 Frank Niessink (VU)
Perspectives on Improving Software Maintenance

2000-02 Koen Holtman (TUE)
Prototyping of CMS Storage Management

2000-03 Carolien M.T. Metselaar (UVA)
Sociaal-Organisatorische Gevolgen van Kennistechnologie;
een Procesbenadering en Actorperspectief

2000-04 Geert de Haan (VU)
ETAG, A Formal Model of Competence Knowledge for User Interface Design

2000-05 Ruud van der Pol (UM)
Knowledge-Based Query Formulation in Information Retrieval

2000-06 Rogier van Eijk (UU)
Programming Languages for Agent Communication

2000-07 Niels Peek (UU)
Decision-theoretic Planning of Clinical Patient Management

2000-08 Veerle Coup (EUR)
Sensitivity Analysis of Decision-Theoretic Networks

2000-09 Florian Waas (CWI)
Principles of Probabilistic Query Optimization

2000-10 Niels Nes (CWI)
Image Database Management System Design Considerations, Algorithms and Architecture

2000-11 Jonas Karlsson (CWI)
Scalable Distributed Data Structures for Database Management

====
2001
====

2001-01 Silja Renooij (UU)
Qualitative Approaches to Quantifying Probabilistic Networks

2001-02 Koen Hindriks (UU)
Agent Programming Languages: Programming with Mental Models

2001-03 Maarten van Someren (UvA)
Learning as Problem Solving

2001-04 Evgueni Smirnov (UM)
Conjunctive and Disjunctive Version Spaces with Instance-Based Boundary Sets

2001-05 Jacco van Ossenbruggen (VU)
Processing Structured Hypermedia: A Matter of Style

2001-06 Martijn van Welie (VU)
Task-Based User Interface Design

2001-07 Bastiaan Schonhage (VU)
Diva: Architectural Perspectives on Information Visualization

2001-08 Pascal van Eck (VU)
A Compositional Semantic Structure for Multi-Agent Systems Dynamics

2001-09 Pieter Jan 't Hoen (RUL)
Towards Distributed Development of Large Object-Oriented Models, Views of Packages as Classes

2001-10 Maarten Sierhuis (UvA)
Modeling and Simulating Work Practice BRAHMS: a Multiagent Modeling and Simulation Language for Work Practice Analysis and Design

2001-11 Tom M. van Engers (VUA)
Knowledge Management:
The Role of Mental Models in Business Systems Design

====
2002
====

2002-01 Nico Lassing (VU)
Architecture-Level Modifiability Analysis

2002-02 Roelof van Zwol (UT)
Modelling and Searching Web-Based Document Collections

2002-03 Henk Ernst Blok (UT)
Database Optimization Aspects for Information Retrieval

2002-04 Juan Roberto Castelo Valdueza (UU)
The Discrete Acyclic Digraph Markov Model in Data Mining

2002-05 Radu Serban (VU)
The Private Cyberspace Modeling Electronic Environments inhabited by Privacy-Concerned Agents

2002-06 Laurens Mommers (UL)
Applied Legal Epistemology;
Building a Knowledge-Based Ontology of the Legal Domain

2002-07 Peter Boncz (CWI)
Monet: A Next-Generation DBMS Kernel for Query-Intensive Applications

2002-08 Jaap Gordijn (VU)
Value Based Requirements Engineering: Exploring Innovative E-Commerce Ideas

2002-09 Willem-Jan van den Heuvel(KUB)
Integrating Modern Business Applications with Objectified Legacy Systems

2002-10 Brian Sheppard (UM)
Towards Perfect Play of Scrabble

2002-11 Wouter C.A. Wijngaards (VU)
Agent Based Modelling of Dynamics: Biological and Organisational Applications

2002-12 Albrecht Schmidt (Uva)
Processing XML in Database Systems

2002-13 Hongjing Wu (TUE)
A Reference Architecture for Adaptive Hypermedia Applications

2002-14 Wieke de Vries (UU)
Agent Interaction: Abstract Approaches to Modelling, Programming and Verifying Multi-Agent Systems

2002-15 Rik Eshuis (UT)
Semantics and Verification of UML Activity Diagrams for Workflow Modelling

2002-16 Pieter van Langen (VU)
The Anatomy of Design: Foundations, Models and Applications

2002-17 Stefan Manegold (UVA)
Understanding, Modeling, and Improving Main-Memory Database Performance

====
2003
====

2003-01 Heiner Stuckenschmidt (VU)
Ontology-Based Information Sharing in Weakly Structured Environments

2003-02 Jan Broersen (VU)
Modal Action Logics for Reasoning About Reactive Systems

2003-03 Martijn Schuemie (TUD)
Human-Computer Interaction and Presence in Virtual Reality Exposure Therapy

2003-04 Milan Petkovic (UT)
Content-Based Video Retrieval Supported by Database Technology

2003-05 Jos Lehmann (UVA)
Causation in Artificial Intelligence and Law - A Modelling Approach

2003-06 Boris van Schooten (UT)
Development and Specification of Virtual Environments

2003-07 Machiel Jansen (UvA)
Formal Explorations of Knowledge Intensive Tasks

2003-08 Yongping Ran (UM)
Repair Based Scheduling

2003-09 Rens Kortmann (UM)
The Resolution of Visually guided Behaviour

2003-10 Andreas Lincke (UvT)
Electronic Business Negotiation: Some Experimental Studies on the Interaction between Medium, Innovation Context and Culture

2003-11 Simon Keizer (UT)
Reasoning under Uncertainty in Natural Language Dialogue using Bayesian Networks

2003-12 Roeland Ordelman (UT)
Dutch Speech Recognition in Multimedia Information Retrieval

2003-13 Jeroen Donkers (UM)
Nosce Hostem - Searching with Opponent Models

2003-14 Stijn Hoppenbrouwers (KUN)
Freezing Language: Conceptualisation Processes across ICT-Supported Organisations

2003-15 Mathijs de Weerdt (TUD)
Plan Merging in Multi-Agent Systems

2003-16 Menzo Windhouwer (CWI)
Feature Grammar Systems - Incremental Maintenance of Indexes to Digital Media Warehouses

2003-17 David Jansen (UT)
Extensions of Statecharts with Probability, Time, and Stochastic Timing

2003-18 Levente Kocsis (UM)
Learning Search Decisions

====
2004
====

2004-01 Virginia Dignum (UU)
A Model for Organizational Interaction: Based on Agents, Founded in Logic

2004-02 Lai Xu (UvT)
Monitoring Multi-party Contracts for E-business

2004-03 Perry Groot (VU)
A Theoretical and Empirical Analysis of Approximation in Symbolic Problem Solving

2004-04 Chris van Aart (UVA)
Organizational Principles for Multi-Agent Architectures

2004-05 Viara Popova (EUR)
Knowledge Discovery and Monotonicity

2004-06 Bart-Jan Hommes (TUD)
The Evaluation of Business Process Modeling Techniques

2004-07 Elise Boltjes (UM)
Voorbeeldig Onderwijs; Voorbeeldgestuurd Onderwijs, een Opstap naar Abstract Denken, vooral voor Meisjes

2004-08 Joop Verbeek(UM)
Politie en de Nieuwe Internationale Informatiemarkt, Grensregionale Politiële Gegevensuitwisseling en Digitale Expertise

2004-09 Martin Caminada (VU)
For the Sake of the Argument; Explorations into Argument-Based Reasoning

2004-10 Suzanne Kabel (UVA)
Knowledge-Rich Indexing of Learning-Objects

2004-11 Michel Klein (VU)
Change Management for Distributed Ontologies

2004-12 The Duy Bui (UT)
Creating Emotions and Facial Expressions for Embodied Agents

2004-13 Wojciech Jamroga (UT)
Using Multiple Models of Reality: On Agents who Know how to Play

2004-14 Paul Harrenstein (UU)
Logic in Conflict. Logical Explorations in Strategic Equilibrium

2004-15 Arno Knobbe (UU)
Multi-Relational Data Mining

2004-16 Federico Divina (VU)
Hybrid Genetic Relational Search for Inductive Learning

2004-17 Mark Winands (UM)
Informed Search in Complex Games

2004-18 Vania Bessa Machado (UvA)
Supporting the Construction of Qualitative Knowledge Models

2004-19 Thijs Westerveld (UT)
Using Generative Probabilistic Models for Multimedia Retrieval

2004-20 Madelon Evers (Nyenrode)
Learning from Design: Facilitating Multidisciplinary Design Teams

====
2005
====

2005-01 Floor Verdenius (UVA)
Methodological Aspects of Designing Induction-Based Applications

2005-02 Erik van der Werf (UM))
AI Techniques for the Game of Go

2005-03 Franc Grootjen (RUN)
A Pragmatic Approach to the Conceptualisation of Language

2005-04 Nirvana Meratnia (UT)
Towards Database Support for Moving Object data

2005-05 Gabriel Infante-Lopez (UVA)
Two-Level Probabilistic Grammars for Natural Language Parsing

2005-06 Pieter Spronck (UM)
Adaptive Game AI

2005-07 Flavius Frasincar (TUE)
Hypermedia Presentation Generation for Semantic Web Information Systems

2005-08 Richard Vdovjak (TUE)
A Model-Driven Approach for Building Distributed Ontology-Based Web Applications

2005-09 Jeen Broekstra (VU)
Storage, Querying and Inferencing for Semantic Web Languages

2005-10 Anders Bouwer (UVA)
Explaining Behaviour: Using Qualitative Simulation in Interactive Learning Environments

2005-11 Elth Ogston (VU)
Agent Based Matchmaking and Clustering - A Decentralized Approach to Search

2005-12 Csaba Boer (EUR)
Distributed Simulation in Industry

2005-13 Fred Hamburg (UL)
Een Computermodel voor het Ondersteunen van Euthanasiebeslissingen

2005-14 Borys Omelayenko (VU)
Web-Service Configuration on the Semantic Web; Exploring how Semantics meets Pragmatics

2005-15 Tibor Bosse (VU)
Analysis of the Dynamics of Cognitive Processes

2005-16 Joris Graaumans (UU)
Usability of XML Query Languages

2005-17 Boris Shishkov (TUD)
Software Specification Based on Re-usable Business Components

2005-18 Danielle Sent (UU)
Test-selection Strategies for Probabilistic Networks

2005-19 Michel van Dartel (UM)
Situated Representation

2005-20 Cristina Coteanu (UL)
Cyber Consumer Law, State of the Art and Perspectives

2005-21 Wijnand Derks (UT)
Improving Concurrency and Recovery in Database Systems by Exploiting Application Semantics

====
2006
====

2006-01 Samuil Angelov (TUE)
Foundations of B2B Electronic Contracting

2006-02 Cristina Chisalita (VU)
Contextual Issues in the Design and use of Information Technology in Organizations

2006-03 Noor Christoph (UVA)
The Role of Metacognitive Skills in Learning to Solve Problems

2006-04 Marta Sabou (VU)
Building Web Service Ontologies

2006-05 Cees Pierik (UU)
Validation Techniques for Object-Oriented Proof Outlines

2006-06 Ziv Baida (VU)
Software-aided Service Bundling - Intelligent Methods & Tools for Graphical Service Modeling

2006-07 Marko Smiljanic (UT)
XML Schema Matching – Balancing Efficiency and Effectiveness by means of Clustering

2006-08 Eelco Herder (UT)
Forward, Back and Home Again - Analyzing User Behavior on the Web

2006-09 Mohamed Wahdan (UM)
Automatic Formulation of the Auditor's Opinion

2006-10 Ronny Siebes (VU)
Semantic Routing in Peer-to-Peer Systems

2006-11 Joeri van Ruth (UT)
Flattening Queries over Nested Data Types

2006-12 Bert Bongers (VU)
Interactivation - Towards an E-cology of People, our Technological Environment, and the Arts

2006-13 Henk-Jan Lebbink (UU)
Dialogue and Decision Games for Information Exchanging Agents

2006-14 Johan Hoorn (VU)
Software Requirements: Update, Upgrade, Redesign - towards a Theory of Requirements Change

2006-15 Rainer Malik (UU)
CONAN: Text Mining in the Biomedical Domain

2006-16 Carsten Riggelsen (UU)
Approximation Methods for Efficient Learning of Bayesian Networks

2006-17 Stacey Nagata (UU)
User Assistance for Multitasking with Interruptions on a Mobile Device

2006-18 Valentin Zhizhkun (UVA)
Graph transformation for Natural Language Processing

2006-19 Birna van Riemsdijk (UU)
Cognitive Agent Programming: A Semantic Approach

2006-20 Marina Velikova (UvT)
Monotone Models for Prediction in Data Mining

2006-21 Bas van Gils (RUN)
Aptness on the Web

2006-22 Paul de Vrieze (RUN)
Fundaments of Adaptive Personalisation

2006-23 Ion Juvina (UU)
Development of Cognitive Model for Navigating on the Web

2006-24 Laura Hollink (VU)
Semantic Annotation for Retrieval of Visual Resources

2006-25 Madalina Drugan (UU)
Conditional Log-likelihood MDL and Evolutionary MCMC

2006-26 Vojkan Mihajlovic (UT)
Score Region Algebra: A Flexible Framework for Structured Information Retrieval

2006-27 Stefano Bocconi (CWI)
Vox Populi: Generating Video Documentaries from Semantically Annotated Media Repositories

2006-28 Borkur Sigurbjornsson (UVA)
Focused Information Access using XML Element Retrieval

====
2007
====

2007-01 Kees Leune (UvT)
Access Control and Service-Oriented Architectures

2007-02 Wouter Teepe (RUG)
Reconciling Information Exchange and Confidentiality: A Formal Approach

2007-03 Peter Mika (VU)
Social Networks and the Semantic Web

2007-04 Jurriaan van Diggelen (UU)
Achieving Semantic Interoperability in Multi-agent Systems: a Dialogue-Based Approach

2007-05 Bart Schermer (UL)
Software Agents, Surveillance, and the Right to Privacy: a Legislative Framework for Agent-enabled Surveillance

2007-06 Gilad Mishne (UVA)
Applied Text Analytics for Blogs

2007-07 Natasa Jovanovic' (UT)
To Whom It May Concern - Addressee Identification in Face-to-Face Meetings

2007-08 Mark Hoogendoorn (VU)
Modeling of Change in Multi-Agent Organizations

2007-09 David Mobach (VU)
Agent-Based Mediated Service Negotiation

2007-10 Huib Aldewereld (UU)
Autonomy vs. Conformity: an Institutional Perspective on Norms and Protocols

2007-11 Natalia Stash (TUE)
Incorporating Cognitive/Learning Styles in a General-Purpose Adaptive Hypermedia System

2007-12 Marcel van Gerven (RUN)
Bayesian Networks for Clinical Decision Support: A Rational Approach to Dynamic Decision-Making under Uncertainty

2007-13 Rutger Rienks (UT)
Meetings in Smart Environments; Implications of Progressing Technology

2007-14 Niek Bergboer (UM)
Context-Based Image Analysis

2007-15 Joyca Lacroix (UM)
NIM: a Situated Computational Memory Model

2007-16 Davide Grossi (UU)
Designing Invisible Handcuffs. Formal investigations in Institutions and Organizations for Multi-agent Systems

2007-17 Theodore Charitos (UU)
Reasoning with Dynamic Networks in Practice

2007-18 Bart Orriens (UvT)
On the Development an Management of Adaptive Business Collaborations

2007-19 David Levy (UM)
Intimate Relationships with Artificial Partners

2007-20 Slinger Jansen (UU)
Customer Configuration Updating in a Software Supply Network

2007-21 Karianne Vermaas (UU)
Fast Diffusion and Broadening use: A Research on Residential Adoption and Usage of Broadband Internet in the Netherlands between 2001 and 2005

2007-22 Zlatko Zlatev (UT)
Goal-Oriented Design of Value and Process Models from Patterns

2007-23 Peter Barna (TUE)
Specification of Application Logic in Web Information Systems

2007-24 Georgina Ramrez Camps (CWI)
Structural Features in XML Retrieval

2007-25 Joost Schalken (VU)
Empirical Investigations in Software Process Improvement

====
2008
====

2008-01 Katalin Boer-Sorban (EUR)
Agent-Based Simulation of Financial Markets: A Modular, Continuous-Time approach

2008-02 Alexei Sharpanskykh (VU)
On Computer-Aided Methods for Modeling and Analysis of Organizations

2008-03 Vera Hollink (UVA)
Optimizing Hierarchical Menus: a Usage-Based Approach

2008-04 Ander de Keijzer (UT)
Management of Uncertain Data - towards Unattended Integration

2008-05 Bela Mutschler (UT)
Modeling and Simulating Causal Dependencies on Process-Aware Information Systems from a Cost Perspective

2008-06 Arjen Hommersom (RUN)
On the Application of Formal Methods to Clinical Guidelines, an Artificial Intelligence Perspective

2008-07 Peter van Rosmalen (OU)
Supporting the Tutor in the Design and Support of Adaptive E-learning

2008-08 Janneke Bolt (UU)
Bayesian Networks: Aspects of Approximate Inference

2008-09 Christof van Nimwegen (UU)
The Paradox of the Guided User: Assistance can be Counter-Effective

2008-10 Wauter Bosma (UT)
Discourse Oriented Summarization

2008-11 Vera Kartseva (VU)
Designing Controls for Network Organizations: A Value-Based Approach

2008-12 Jozsef Farkas (RUN)
A Semiotically Oriented Cognitive Model of Knowledge Representation

2008-13 Caterina Carraciolo (UVA)
Topic Driven Access to Scientific Handbooks

2008-14 Arthur van Bunningen (UT)
Context-Aware Querying; Better Answers with Less Effort

2008-15 Martijn van Otterlo (UT)
The Logic of Adaptive Behavior: Knowledge Representation and Algorithms for the Markov Decision Process Framework in First-Order Domains

2008-16 Henriette van Vugt (VU)
Embodied Agents from a User's Perspective

2008-17 Martin Op 't Land (TUD)
Applying Architecture and Ontology to the Splitting and Allying of Enterprises

2008-18 Guido de Croon (UM)
Adaptive Active Vision

2008-19 Henning Rode (UT)
From Document to Entity Retrieval: Improving Precision and Performance of Focused Text Search

2008-20 Rex Arendsen (UVA)
Geen Bericht, Goed Bericht. Een Onderzoek naar de Effecten van de Introductie van Elektronisch Berichtenverkeer met de Overheid op de Administratieve Lasten van Bedrijven

2008-21 Krisztian Balog (UVA)
People Search in the Enterprise

2008-22 Henk Koning (UU)
Communication of IT-Architecture

2008-23 Stefan Visscher (UU)
Bayesian Network Models for the Management of Ventilator-Associated Pneumonia

2008-24 Zharko Aleksovski (VU)
Using Background Knowledge in Ontology Matching

2008-25 Geert Jonker (UU)
Efficient and Equitable Exchange in Air Traffic Management Plan Repair using Spender-Signed Currency

2008-26 Marijn Huijbregts (UT)
Segmentation, Diarization and Speech Transcription: Surprise Data Unraveled

2008-27 Hubert Vogten (OU)
Design and Implementation Strategies for IMS Learning Design

2008-28 Ildiko Flesch (RUN)
On the Use of Independence Relations in Bayesian Networks

2008-29 Dennis Reidsma (UT)
Annotations and Subjective Machines - of Annotators, Embodied Agents, Users, and Other Humans

2008-30 Wouter van Atteveldt (VU)
Semantic Network Analysis: Techniques for Extracting, Representing and Querying Media Content

2008-31 Loes Braun (UM)
Pro-Active Medical Information Retrieval

2008-32 Trung H. Bui (UT)
Toward Affective Dialogue Management using Partially Observable Markov Decision Processes

2008-33 Frank Terpstra (UVA)
Scientific Workflow Design; Theoretical and Practical Issues